INITIAL MEMBER SELECTION AND COVARIANCE LOCALIZATION STUDY OF ENSEMBLE KALMAN FILTER BASED DATA ASSIMILATION

A Thesis

by

YEUNG YIP

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2011

Major Subject: Petroleum Engineering

INITIAL MEMBER SELECTION AND COVARIANCE LOCALIZATION STUDY OF ENSEMBLE KALMAN FILTER BASED DATA ASSIMILATION

A Thesis

by

YEUNG YIP

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee,	Akhil Datta-Gupta
Committee Members,	Gioia Falcone
	Yalchin Efendiev
Head of Department,	Stephan A. Holditch

May 2011

Major Subject: Petroleum Engineering

ABSTRACT

Initial Member Selection and Covariance Localization

Study of Ensemble Kalman Filter Based Data Assimilation. (May 2011)

Yeung Yip, B.S., Texas A&M University

Chair of Advisory Committee: Dr. Akhil Datta-Gupta

Petroleum engineers generate reservoir simulation models to optimize production and maximize recovery. History matching is one of the methods used to calibrate the reservoir models. During traditional history matching, individual model variable parameters (permeability, relative permeability, initial water saturation, etc) are adjusted until the production history is matched using the updated reservoir model. However, this method of utilizing only one model does not help capture the full range of system uncertainty. Another drawback is that the entire model has to be matched from the initial time when matching for new observation data.

Ensemble Kalman Filter (EnKF) is a data assimilation technique that has gained increasing interest in the application of petroleum history matching in recent years. The basic methodology of the EnKF consists of the forecast step and the update step. This data assimilation method utilizes a collection of state vectors, known as an ensemble, which are simulated forward in time.

In other words, each ensemble member represents a reservoir model (realization). Subsequently, during the update step, the sample covariance is computed from the ensemble, while the collection of state vectors is updated using the formulations which involve this updated sample covariance.

When a small ensemble size is used for a large, field-scale model, poor estimate of the covariance matrix could occur (Anderson and Anderson 1999; Devegowda and Arroyo 2006). To mitigate such problem, various covariance conditioning schemes have been proposed to improve the performance of EnKF, without the use of large ensemble sizes that require enormous computational resources.

In this study, we implemented EnKF coupled with these various covariance localization schemes: Distance-based, Streamline trajectory-based, and Streamline sensitivity-based localization and Hierarchical EnKF on a synthetic reservoir field case study. We will describe the methodology of each of the covariance localization schemes with their characteristics and limitations.

ACKNOWLEDGEMENTS

I would like to acknowledge my academic advisor, Dr Akhil Datta-Gupta for his financial and academic support during the course of my master's degree program. I would also like to acknowledge Dr. Falcone and Dr. Efendiev for being my committee members. Finally, I would like to thank former MCERI students: Deepak Devegowda, Elkin Arroyo-Negrete, Ajitabh Kumar, Akella Santha, Eduardo Jimenez ,Adedayo Oyerinde, Ahmed Al-Huthali, Jong-Uk Kim, and Qing Tao for their contributions and achievements in our research area. Also to current students: Shingo Watanabe, Alvaro Jose Rey Amaya, Jichao Yin, Eric Bhark, , Song Du, Han-Young Park, Baljit Sehbi, Satyajit Taware, Shusei Tanaka and Suksang Kang, for their friendship and inspirations.

NOMENCLATURE

$\mathbf{C}$	Connectivity matrix
$\mathbf{C_D}$	Data covariance matrix
$\mathbf{C}_{M^s,d}$	Cross-covariance matrix between data and model parameters
$\mathbf{C}_K^P$	Updated model covariance matrix
$\mathbf{D}$	Diagonal matrix of eigenvalues of the matrix F
$\mathbf{D}_{obs,k}$	Ensemble of the observation data
$\mathbf{d}_{obs}$	Observation data vector
$\mathbf{d}_k^{cal}$	Calculated or theoretical observation vector
$\mathbf{d}_{true}$	True observation data vector
E	Connectivity matrix with the mean vector subtracted
F	Covariance matrix of matrix E
ε_k	Measurement error
$\mathbf{H}$	Measurement matrix
K	Absolute Permeability
K_k	Kalman Gain
$log(K)$	Log of absolute permeability
k_r	Relative permeability
$\mathbf{m}_k^{sta}$	Vector of static model variables
$\mathbf{m}_k^{dyn}$	Vector of dynamic model variables
N_d	Number of observation data

N_e	Number of ensemble members
$\boldsymbol{\rho}$	Covariance localizing function
$\mathbf{R}^N$	N numbers of real numbers
t	Time
τ	Time of flight
y_k	Model state vector
y_{N_{states},N_e}	Augmented Model state vector
x_i	Vector incorporated into calculating Euclidean distance
$\delta_{i,j}$	Euclidean distance
V	Matrix of eigenvectors

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

Page

Page

Page

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Background of Ensemble Kalman Filter (EnKF)

Petroleum engineers generate reservoir simulation models to optimize production and maximize recovery. History matching is one of the methods used to calibrate the reservoir models. During traditional history matching, individual model variable parameters (permeability, relative permeability, initial water saturation, etc) are adjusted until the production history is matched using the updated reservoir model. However, this method of utilizing only one model does not help capture the full range of system uncertainty. Another drawback is that the entire model has to be matched from the initial time when matching for new observation data.

Ensemble Kalman Filter (EnKF) is a data assimilation technique that has gained increasing interest in the application of petroleum history matching in recent years. The basic methodology of the EnKF consists of the forecast step and the update step. This data assimilation method utilizes a collection of state vectors, known as an ensemble, which are simulated forward in time. In other words, each ensemble member represents a reservoir model (realization). Subsequently, during the update step, the sample covariance is computed from the ensemble, while the collection of state vectors is updated using the formulations which involve this updated sample covariance.

This thesis follows the style of *SPE Journal*.

A key component of the EnKF statistical formulations is the Kalman Gain which contains the cross-covariance between the model variables and calculated responses. Since the EnKF updating is highly sensitive to the Kalman Gain, it is expected that a poor estimate of the cross-covariance matrix could severely degrade EnKF performance (Anderson and Anderson, 1999; Furrer and Bengtsson, 2004).

When a small ensemble size is used for a large, field-scale model, poor estimate of the covariance matrix could occur. To mitigate such problem, various covariance conditioning schemes have been proposed to improve the performance of EnKF, without the use of large ensemble sizes that require enormous computational resources.

In this study, we implemented EnKF coupled with these various covariance localization schemes: Distance-based, Streamline trajectory-based, and Streamline sensitivity-based localization and Hierarchical EnKF on a synthetic reservoir field case study. We will describe the methodology of each of the covariance localization schemes with their characteristics and limitations.

We hope to obtain insights for choosing the suitable covariance localization schemes, aided by the knowledge of their limitations, for future application of EnKF to reservoir characterization.

Objectives of Initial Member Selection

It is imperative that a small ensemble of realizations is used in order to reduce the computational requirements for the implementation of EnKF on a large-scale model. However, it is also important to minimize the errors associated with the use of small ensemble sizes. All these contribute to the motivation to experiment with various algorithms to select the key initial members for a large-scale field model, as a pre-processing step to aid EnKF data assimilation.

Objectives of Covariance Localization Study

Covariance localization basically means localizing the effect of an observation to the state variables that are 'closer' to the observations. The various localization methods proposed in the literature have the common goal of removing the spurious terms in the cross-covariance matrix, this matrix is in turn used to update the state vectors during the EnKF update process. This is done by conditioning the Kalman Gain through a localizing function. Each localization scheme distinguishes itself from each other by how this localizing function, also known as the Schur product or the multiplier function, is computed.

The motivation of covariance localization is to achieve a similar level of EnKF performance if a larger ensemble size would have been used. However, since it requires enormous computational resources to perform history matching if the ensemble size is large. Various cases of EnKF with and without localization are implemented on a highly heterogeneous synthetic field case. In this thesis, we will describe the methodology of

each localization scheme with their characteristics and limitations on this kind of case setting. To judge the effectiveness of each covariance localization method, the quality of the dynamic response history match and the performance of parameter estimation with respect to the reference model are the main parameters examined.

Thesis Outline

In this chapter, we make a general description of the Ensemble Kalman Filter (EnKF). We discuss the motivation for the process of initial member selection and covariance localization while using the EnKF methodology.

In Chapter II, we discuss the mathematical formulations of the EnKF in greater details.

In Chapter III, we discuss the results of the initial member selection process for the EnKF implementation. We also compare the history matching efficiency of the different initial member selection schemes.

In Chapter IV, we apply the different EnKF covariance localization schemes to two synthetic reservoir models. We also compare the history matching quality of the different covariance localization schemes.

In Chapter V, we conclude and make recommendations for future work.

CHAPTER II

MATHEMATICAL FORMULATIONS FOR THE ENSEMBLE KALMAN FILTER

Mathematical Formulations for Ensemble Kalman Filter

In the context of petroleum reservoir history matching, Ensemble Kalman Filter (EnKF) is a data assimilation technique which assimilates observation information to a numerical model, typically a reservoir simulator. Ensemble Kalman Filter has been diversely applied since its implementation in 2002. In this chapter only the statistical details of the EnKF in the context of reservoir characterization will be discussed.

In a nutshell, in the application of reservoir modeling, Ensemble Kalman Filter forms the best estimate of the model states and parameters that honor the observations up to the current time. It should be foremost noted that the EnKF is optimal for Gaussian models and linear model systems (Devegowda, Arroyo-Negrate et al. 2009). In other words, the underlying non-linearity of typical multiphase flow simulation systems, can degrade the EnKF update process leading to geo-statistically less realistic changes to the model realizations.

[1]Mathematical derivations discussed in the following can be referred to Reference Evensen G. Data Assimilation: The Ensemble Kalman Filter. Springer, 2006 p.13~p.42.

The basic methodology of EnKF consists of the forecast step and the update step. The forecast step is to advance the state vectors of the ensemble members from a data assimilation time to the next. The ensemble members are sampled from a suite of reservoir models. Through the data assimilation time, the advancement of the state vectors can be achieved by using a reservoir simulator (FrontSim developed by Schlumberger is used in this study).

Subsequently, during the update step, the covariance matrix is computed explicitly which is typical during the Kalman Filter process. However, during EnKF, the ensemble covariance matrix can be obtained by the propagated ensemble replicates of state vectors. In turn, incorporating this covariance matrix into the Kalman Gain computation, the state vectors are updated. Next the forecast step described earlier is repeated for the next data assimilation time until the most current time.

Before illustrating the statistical formulations of EnKF, we first need to clarify the components of the state vector. The state vector for each ensemble member is comprised of model parameters, also known as static model variables m_k^{sta}, dynamic model variables m_k^{dyn} and the calculated observable data d_k^{cal} as outlined in the following:

$$y_k = \begin{Bmatrix} m_k^{sta} \\ m_k^{dyn} \\ d_k^{cal} \end{Bmatrix} \tag{2.2}$$

The state vector denoted as y_k typically includes static variables (e.g. porosity, permeability), dynamic variables (e.g. pressure and phase saturations), and observations

such as water-cut (WWCT), bottom-hole pressure (WBHP) and gas-oil ratio (GOR), etc. at each simulation grid block.

In one of the synthetic field cases used in this thesis, the state vector is comprised of log permeability, pressure, saturation, water-cut and bottom-hole pressure. Another synthetic case is comprised of the same elements but with the absence of the bottom-hole pressure. For the former case, the augmented state vector is as follows:

$$y_{N_{states},N_e} = \begin{Bmatrix} \log k_{1,1} \cdots \log k_{1,N_e} \\ \vdots \quad\quad \vdots \\ \log k_{N,1} \cdots \log k_{N,N_e} \\ P_{1,1} \cdots P_{1,N_e} \\ \vdots \quad\quad \vdots \\ P_{N,1} \cdots P_{N,N_e} \\ S_{1,1} \cdots S_{1,N_e} \\ \vdots \quad\quad \vdots \\ S_{N,1} \cdots S_{N,N_e} \\ WWCT_{1,1} \cdots WWCT_{1,N_e} \\ \vdots \quad\quad \vdots \\ WWCT_{N_d,1} \cdots WWCT_{N_d,N_e} \\ WBHP_{1,1} \cdots WBHP_{1,N_e} \\ \vdots \quad\quad \vdots \\ WBHP_{N_d,1} \cdots WBHP_{N_d,N_e} \end{Bmatrix} \tag{2.3}$$

The state vector y_k has a dimension of (N_{states} x N_e) where N_{states} is the size of the state vector, the number of matrix rows outlined in Eq.2.3. N_e is the number of ensemble

members, the number of matrix columns. Within the matrix, *N* denotes the number of grid blocks, while N_d denotes the number of observations.

The following section will explain the components of the ENKF statistical formulations.

Observed Data

The observed data typically involves measurement errors because measurements are typically subject to uncertainty. Assume that some measurements are obtained at the time step k,

$$d_{obs,k} = d_{true} + \varepsilon_k = Hy_k + \varepsilon_k \tag{2.4}$$

The observable data denoted as $d_{obs,k}$ is actually the perturbed observation vector from the true data (if noise did not exist), since physical measurements always involve noise. Since only synthetic field cases are studied in this work, no actual observation is obtained using any sort of measurement device. Therefore, the actual observation d_{true} referred to in this study is the output of running the forward simulation using the 'true' realization. The 'true' realization is arbitrarily chosen among the initial realizations. The initial realizations were created using Sequential Gaussian Simulation.

Referring to Eq. 2.4 again, ε_k is the unknown measurement error at time step *k*. The actual value of ε_k is not known. Therefore, the following assumptions are made:

i) It is a Gaussian distribution with a dimension of N_d x1.

ii) Its mean is zero, i.e. $\overline{\varepsilon}_k = 0$.

iii) Its covariance matrix is defined as

$$C_{D,k} = E[\varepsilon_k \varepsilon_k^T] = \overline{\varepsilon_k \varepsilon_k}^T \tag{2.5}$$

with dimensions equal to N_d x N_d where as outlined earlier N_d is the number of total observations. $C_{D,k}$ is known as the observation error covariance matrix. We assume that the errors in the observation are not correlated, therefore $C_{D,k}$ is simply a diagonal matrix. In summary, $d_{obs,k}$ is the perturbed observation vector of each of the ensemble members at time step *k*. It is obtained by adding random noise ε_k to the "true" observation d_{true} (Burgers, Leeuwen et al. 1998). The perturbed observations $d_{obs,k}$ of each ensemble member from Eq.2.4 are incorporated into the following equation in order to update the state vectors:

$$\mathbf{y}_k^u = \mathbf{y}_k^P + \mathbf{K}_k\left(\mathbf{d}_{obs,k} - \mathbf{H}_k \mathbf{y}_k^P\right) \tag{2.6}$$

Identity Matrix

H_k is the identity matrix outlined in the following equation Eq.2.7. Also known as the measurement operator, it has a dimension of (N_d x N_{states}). Referring to the above Eq.2.6, H_k basically extracts the computed observation response and gives the linear relationship between the measurements and the states.

$$H_k = [0 \ \mathrm{I}] \tag{2.7}$$

Covariance Matrix of State Vector

The error covariance matrix C_k^P at a certain time step is defined as:

$$C_k^P = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (y_{k,i}^P - y_{k,true}^P)(y_{k,i}^P - y_{k,true}^P)^T \tag{2.8}$$

Since the actual state vector is not known, the actual state vector is instead approximated by the mean of the ensemble state vectors as outlined in the following:

$$C_k^P = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (y_{k,i}^P - \bar{y}_k^P)(y_{k,i}^P - \bar{y}_k^P)^T \tag{2.9}$$

where $\bar{y}_k^p$ is the averaged state vector of all ensemble members:

$$\bar{y}_k^p = \frac{1}{N_e} \sum_{i=1}^{N_e} y_{k,i}^P \tag{2.10}$$

The update error covariance matrix is estimated from the propagated state vectors sampled from the initial suite of realizations. This sample covariance matrix is a highlight feature of Ensemble Kalman Filter in that it is estimated using the updated ensembles of state vectors. The covariance matrix of the state vector is outlined as:

$$\mathbf{C}^P = \begin{pmatrix} \mathbf{C}_{\mathbf{m}^{sta}} & \mathbf{C}_{\mathbf{m}^{sta},\mathbf{m}^{dym}} & \mathbf{C}_{\mathbf{m}^{sta},\mathbf{d}^{cal}} \\ \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} & \mathbf{C}_{\mathbf{m}^{dym}} & \mathbf{C}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} \\ \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{sta},\mathbf{d}^{cal}} & \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} & \mathbf{C}_{\mathbf{d}^{cal}} \end{pmatrix}^p \tag{2.11}$$

The covariance matrix of the static variables $C_{m^{sta}}$ and that of the dynamic variables $C_{m^{dyn}}$ both have a size of $N_m \times N_m$ where N_m is the number of grid blocks. The covariance

matrix of the calculated data $C_{d^{cal}}$ has a size of $N_d \times N_d$ where N_d is the number of measured data. After defining the covariance matrix, the cross covariance matrices $C^P H^T$ and $HC^P H^T$ within the Kalman Gain is defined as follows:

$$\mathbf{C}^P H^T = \begin{pmatrix} \mathbf{C}_{\mathbf{m}^{sta}} & \mathbf{C}_{\mathbf{m}^{sta},\mathbf{m}^{dym}} & \mathbf{C}_{\mathbf{m}^{sta},\mathbf{d}^{cal}} \\ \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} & \mathbf{C}_{\mathbf{m}^{dym}} & \mathbf{C}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} \\ \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{sta},\mathbf{d}^{cal}} & \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} & \mathbf{C}_{\mathbf{d}^{cal}} \end{pmatrix}^p \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} C_{m^{sta},d^{cal}} \\ C_{m^{dym},d^{cal}} \\ C_{d^{cal}} \end{pmatrix}^P \tag{2.12}$$

where

$$H\mathbf{C}^P H^T = C^P_{d^{cal}} \tag{2.13}$$

It is apparent from Eq. 2.12 that we do not have to compute or store the full matrix C^P but multiply it with the operator *H*. We are only computing the relevant matrix elements in C^P, so during EnKF implementation, we compute the following instead of the entire covariance matrix shown earlier in Eq. 2.11:

$$C^P_k H^T = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (y^P_{k,i} - \bar{y}^P_k)[H(y^P_{k,i} - \bar{y}^P_k)]^T \tag{2.14}$$

In addition, $H\mathbf{C}^P H^T$ needs to be computed as follows:

$$HC^P_k H^T = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (Hy^P_{k,i} - H\bar{y}^P_k)[H(y^P_{k,i} - \bar{y}^P_k)]^T \tag{2.15}$$

Kalman Gain

K_k is known as the Kalman Gain matrix with a dimension of (N_{states} *x* N_d). The Kalman Gain K_k is the same for all N_e ensemble members, thus saving computational requirements as opposed to doing N_e independent analysis.

$$K_k = C_k^P H_k^T (H_k C_k^p H_k^T + C_{D,k})^{-1} \tag{2.16}$$

Incorporating the Kalman Gain into the update equation, the following is obtained:

$$\mathbf{y}_k^u = \mathbf{y}_k^P + C_k^P H_k^T (H_k C_k^P H_k^T + C_{D,k})^{-1} \left(\mathbf{d}_{obs,k} - \mathbf{H}_k \mathbf{y}_k^P \right) \tag{2.17}$$

In summary, using the components defined earlier: the matrices $C^P H^T$ and $HC_k^P H^T$ along with the diagonal matrix of expected measurement error variance $C_{D,k}$, we can find the Kalman Gain at each update step.

Summary

The basic methodology of the EnKF consists of the forecast step and the update step. The forecast step is to advance the state vectors of all ensemble members from the initial time step to the next through forward simulation. The purpose of the update step is to update the state vector variables to honor the observed data. The observed data typically involves measurement errors where all measurements are subject to uncertainty. Therefore, in the synthetic field cases discussed in the next chapters, different combinations of measurement error of the observable responses (e.g. water-cut and bottom-hole pressure) are used to test EnKF performance.

In this chapter, we discussed the mathematical basis of EnKF. We would however state some of the observations we consider helpful for the practical application of this methodology. The petroleum reservoir system is highly non-linear, however, we are aware that EnKF is optimal for Gaussian models and linear systems. Therefore, when building the synthetic case, one of our primary concerns is to utilize a suite of individually normally distributed members. All in all, we have to identify the variables that have the most impact on EnKF quality. These variables include the size of the ensemble member and the combinations of measurement errors of the dynamic responses.

CHAPTER III

OPTIMAL INITIAL MEMBER SELECTION OF THE ENSEMBLE KALMAN FILTER

Initial Member Selection

As discussed earlier, it is computationally challenging to use large ensemble sizes during the EnKF implementation especially on a large-scale model or when using more complex compositional simulation. It becomes imperative that a small ensemble is used in order to reduce the computational requirements. However, it is also important to minimize the errors associated with the use of small ensemble sizes. All these contribute to the need to select the key members for EnKF application. This process helps optimize accuracy, computation resources and time.

Before presenting the initial member selection results using the Ng's Spectral Clustering and the Multidimensional Scaling / K-means Clustering algorithms, we will illustrate the need to select the key initial members as a pre-processing process to improve the EnKF implementation. We will briefly describe the various routines used in this part of the thesis. For completeness, we note that there are other clustering methods which include but are not limited to fuzzy c-means clustering and QT-clustering.

Summary of Algorithms Used for Initial Member Selection

We use two algorithms as means to compute the distance/connectivity matrix, and subsequently cluster data points to obtain the key realizations with swept pore volume at

different time-of-flight (TOF) as the input to the algorithms. The first algorithm is the spectral clustering method with specific steps originally proposed by Ng et. Al applied by Ajitabh Kumar on a synthetic water-flood field case, to select key members to use for EnKF later. The second algorithm is a combination of multidimensional scaling and K-means clustering. Both methods use Euclidean distance as the distance measure.

One key aspect of the technique formerly implemented by Ajitabh Kumar on selecting the key initial ensemble members is based on monitoring the dissimilarity among the members. In summary, the dissimilarity distance can be measured by the distance between two realizations that are correlated by the difference in certain responses of the two realizations. In other words, it is not necessary to measure the absolute statistical features of each realization, but rather the "distance" or dissimilarity between them (Scheidt and Caers, 2007).

In the synthetic field case study, the responses concerned are the swept pore volume at certain time-of-flight (TOF). Since flow simulation is a highly non-linear process, selection of members using only static measures such as permeability values is not optimal. Therefore, this technique of selecting members using swept pore volume has an advantage compared to using permeability values alone which contain no information about the flow-related non-linearity. (Scheidt and Caers, 2007).

In summary, the technique used to select the key initial members for EnKF implementation has the following procedures. Fundamentally, the dissimilarities among the ensemble members are based on the evolution of swept pore volume by the injected water though time. The swept pore volume at certain TOF is computed by running the

streamline simulation for all members. Subsequently, the data of swept pore volume at the progressing TOF is recorded. The dissimilarities between the member incremental swept pore volume values at different TOF, are then used as input to the spectral clustering algorithm to select the initial key members. The following section will illustrate the procedures of this specific algorithm.

Spectral Clustering Algorithm

We employed the spectral clustering algorithm proposed by Ng et al., 2002 with specific steps outlined as follows:

Given a set of points: $S = \{s_1,...,s_n\}$ in real space R^l, k subsets of points can be clustered using the following procedures:

i) Construct the affinity matrix $A \in R^{n \times n}$ where $A_{ij} = \exp\left(-\left\|s_i - s_j\right\|^2 / 2\sigma^2\right)$ if $i \neq j$ and $A_{ii} = 0$

ii) Construct the matrix $L = D^{-1/2} A D^{-1/2}$ where D is defined as the diagonal matrix whose (i,i)-element is the sum of A's i-th row. [1]

iii) Find the k largest eigenvectors: $x_1,...,x_k$ of L and form the matrix $X = \left[x_1 x_2 ... x_k\right] \in R^{n \times k}$ by stacking the eigenvectors in columns.

iv) Construct the matrix Y by renormalizing each of X's rows to have unit length such that $Y_{ij} = X_{ij} / \left(\sum_j X_{ij}^2\right)^{1/2}$.

v) Treat each row of Y as a point in R^k and cluster the points into k clusters using K-means (in order to minimize the distortion).

vi) Only assign the original point s_i to cluster j if row i of matrix Y was assigned to cluster j.

In general, a critical step in spectral clustering analysis is the selection of a distance measure. Its fundamental concept is to determine how similar two elements are with each other. In other words, the distance referred to is the dissimilarity between the two elements. Euclidean distance is the distance measure considered in this algorithm.

Multidimensional Scaling & Principal Component Analysis Algorithm

The process of initial member selection described in this section is a variant of the procedure described by Alpak et al., 2008. They utilized a couple of flow-based pattern recognition algorithms to quantify geologic uncertainty of reservoir models in a channelized turbidite reservoir (Alpak et al., 2008).

In this work, we will use a combination of multidimensional scaling (MDS), principal component analysis (PCA) and the K-means algorithm to select an ensemble subset that is representative of a synthetic reservoir case. This algorithm will be referred to as MDS-KMPCA in brief in the following section. The advantages of using this procedure are detailed in the work of Alpak et al., 2008. Before embarking on a detailed description of the workflow, we will describe each of the component parts of this workflow briefly.

Multidimensional Scaling

According to the description in Borg et al., 2005, MDS is a statistical technique used to convert high dimensional to low dimensional data while still retaining the defining characteristics of the dataset. For example, for the case of illustration, let us say we have *I* ensembles of water-cut data, with each ensemble having a total of 160 time steps. We define a connectivity matrix which is a matrix of the distance (in some sense) of all ensembles from each other, that is, a *I* x *I* matrix. Their distance from each other is assumed to be a measure of the dissimilarity of their water-cut responses. An ensemble of vectors of low dimensionality is obtained such that the difference between the norm of the difference of the vectors and the corresponding member of the connectivity matrix is minimized. The connectivity matrix is defined as shown below;

$$C = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix} \tag{3.1}$$

The goal of MDS is, given the connectivity matrix *C,* to find *I* vector

$$x_1, \ldots, x_I \in R^N \tag{3.2}$$

such that

$$\left\| x_i - x_j \right\| = \delta_{i,j} \tag{3.3}$$

for all i,j in I where $\|\cdot\|$ is a vector norm. In classical MDS, this norm is the Euclidean distance. The dimension N is typically chosen to be 2 or 3, where we can plot the vectors x_i to visualize the similarities between the I objects. It is worthwhile to note that the vectors x_i are not unique, where the Euclidean distance may be arbitrarily translated and rotated since these transformations do not change the distances in pairs: $\|x_i - x_j\|$.

Principal Component Analysis

In this thesis, linear principal component analysis (PCA) is a pre-processing step to aid the k-means clustering algorithm in partitioning the data into different clusters. In other words, this step is expected to help the k-means clustering procedure find near-optimal solutions. According to the description in Jolliffe et al., 2002, PCA, also known as Karhunen-Loeve transform (KLT), is a methodology used to transform a matrix of correlated variables into a reduced number of uncorrelated variables. In this procedure, the first principal component is responsible for as much variability in the matrix as possible with each succeeding principal component accounting for most of the rest of the variability.

We describe the steps necessary to compute the prinicipal components in the following lines;

(1) Compute the empirical mean along each dimension of the connectivity matrix C, i=1,...I. Assign the calculated mean values into an empirical mean vector m of dimension I x 1.

(2) Subtract the mean vector *m* from each column of the connectivity matrix, *C* and call it *E*.

(3) Find the co-variance matrix of *E*, and call it *F*. Use the cov command in MATLAB.

(4) Find the matrix *V* of eigenvectors which diagonalizes *F* as shown below;

$$V^{-1} F V = \mathrm{D} \tag{3.4}$$

where *D* above is the diagonal matrix of eigenvalues of *F*.

The comprehensive steps implemented in this algorithm are outlined as the following:

1. Define the dimensions of principal component analysis to be performed, e.g. 2-D or 3-D.

2. Define the number of members to be selected. The number of members to be selected is the number of clusters/centroids.

3. Load/Input dynamic responses, e.g. incremental swept pore volume at progressing time-of-flight.

4. Compute the dot product of the dynamic responses for all time-of-flight of one member and another. Repeat this process until the matrix of connectivity/distances among all members is computed.

5. Diagonalize the connectivity matrix.

6. Perform eigenvalue decomposition to obtain the eigenvectors and eigenvalues of the matrix.

7. Perform principal component analysis. The output variables are known as points X.

8. Perform K-mean cluster analysis of the variables X to obtain the centroids for the number of clusters intended.

The following section describes and discusses the results obtained by using the two algorithms: Spectral Clustering Algorithm; and Multidimensional Scaling & K-means Clustering Algorithm to select initial key members for EnKF.

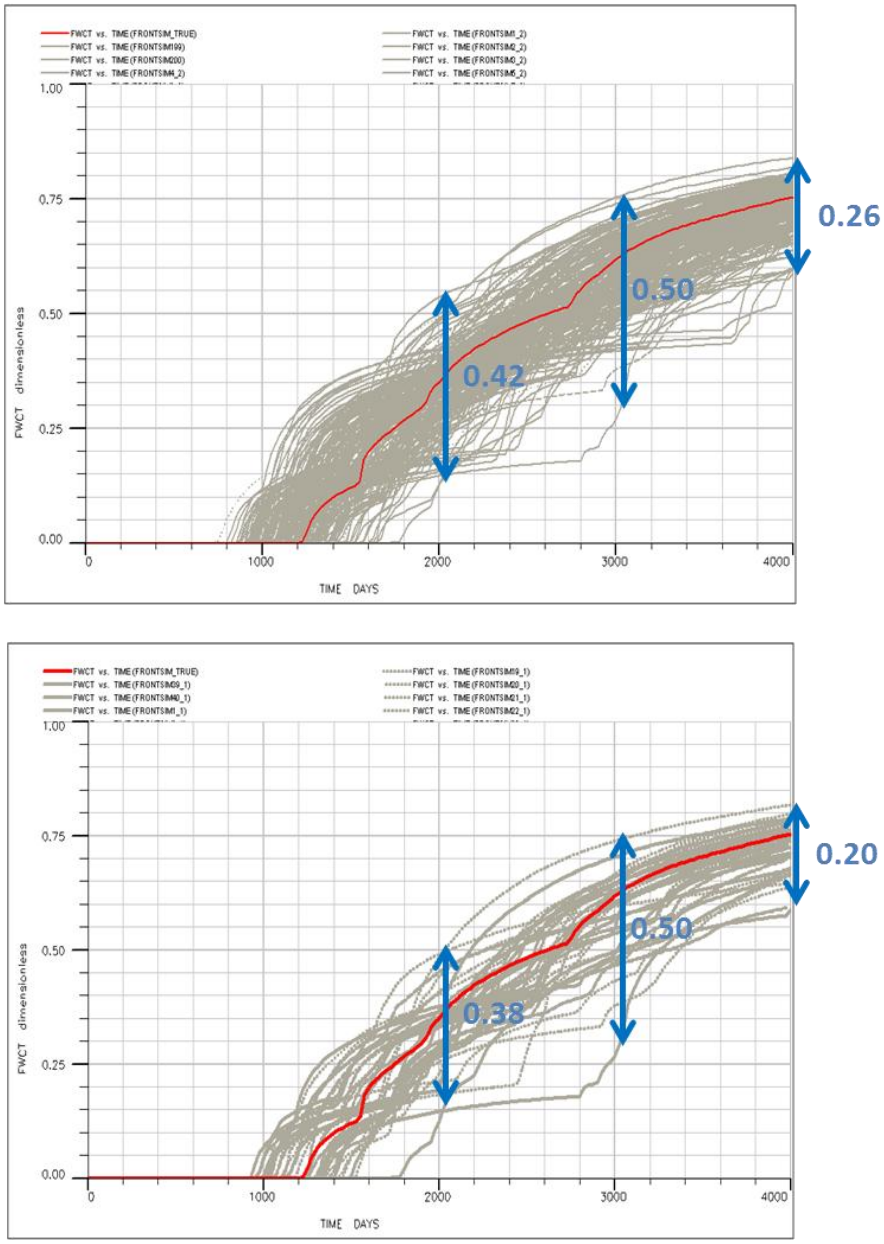


Figure 3.1. The simulated field-wide water cut responses before history-matching for (a) all initial 200 members, (b) 40 members selected by the Spectral Clustering method. Red curve shows the water cut response of the reference model. The vertical blue lines dictate the width of the water cut spread. The numbers next to the lines show the numerical difference between the upper bound and lower bound water cut values at the particular time.

Referring to the results before EnKF data assimilation as shown in Figure 3.1 above and Figure 3.2 below, either key member selection method shows comparable water-cut

spread to the reference case where 200 initial members are used. This suggests that using either technique can achieve a similar quality to that of using a larger pool of realizations.

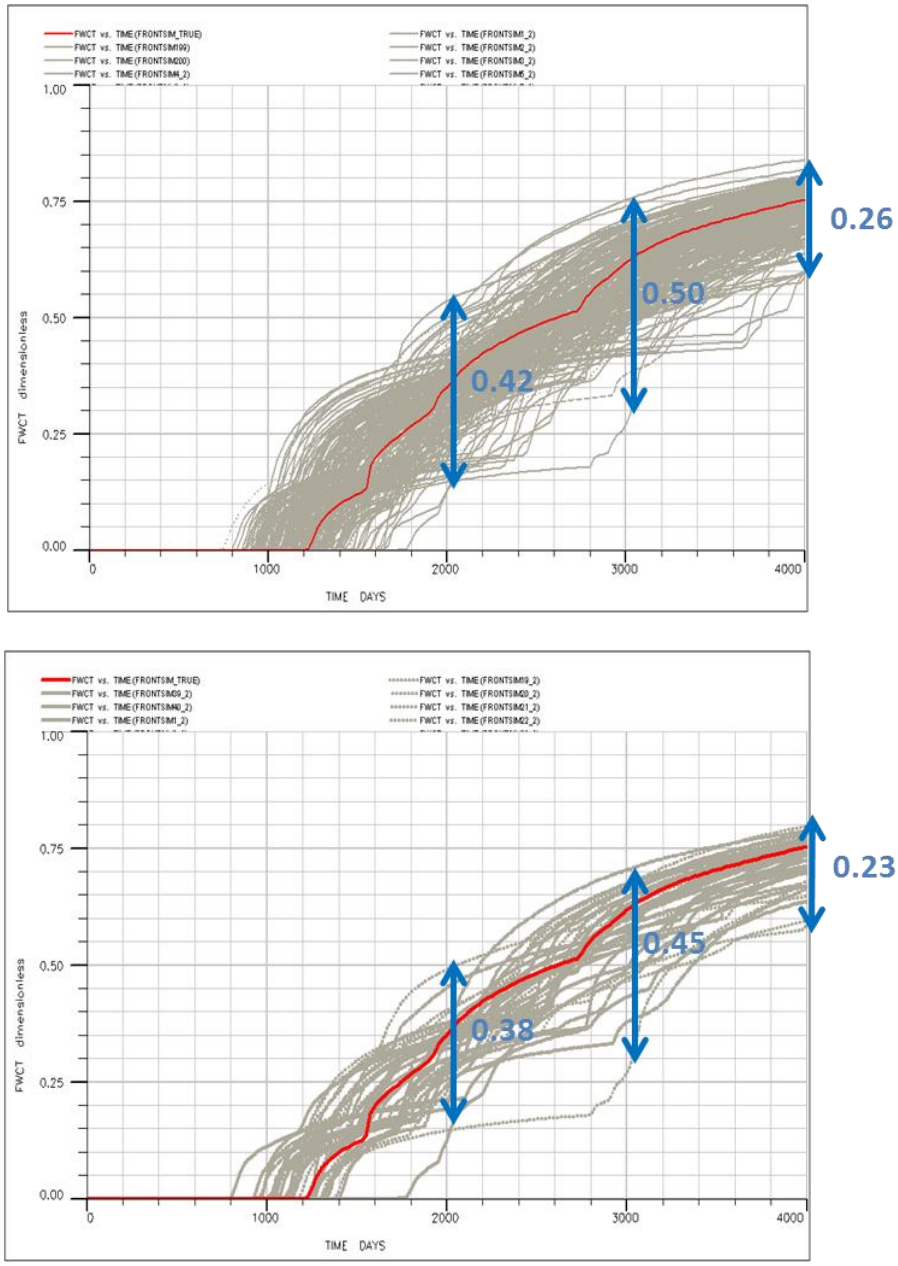


Figure 3.2. The simulated field-wide water cut responses before history-matching for (a) all initial 200 members, (b) 40 members selected by the Multidimensional Scaling & K-means Clustering method. Red curve shows the water cut response of the reference model. The vertical blue lines dictate the width of the water cut spread. The numbers next to the lines show the numerical difference between the upper bound and lower bound water cut values at the particular time.

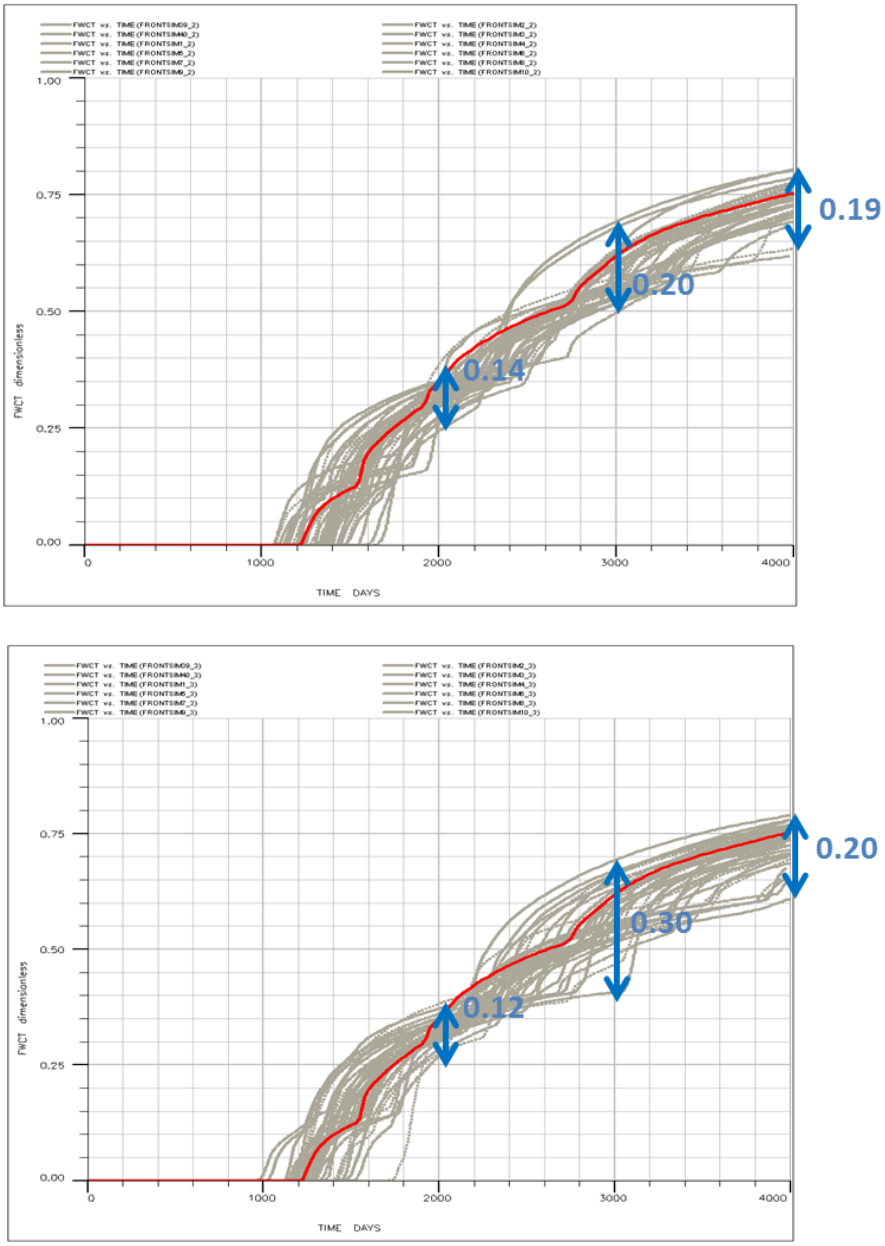


Figure 3.3. The simulated field-wide water cut responses after history-matching for (a) Plain EnKF with 40 members, (b) 40 members selected by the Spectral Clustering method. Red curve shows the water cut response of the reference model. The vertical blue lines dictate the width of the water cut spread. The numbers next to the lines show the numerical difference between the upper bound and lower bound water cut values at the particular time.

Figure 3.3 above and Figure 3.4 below show the history match results after EnKF.

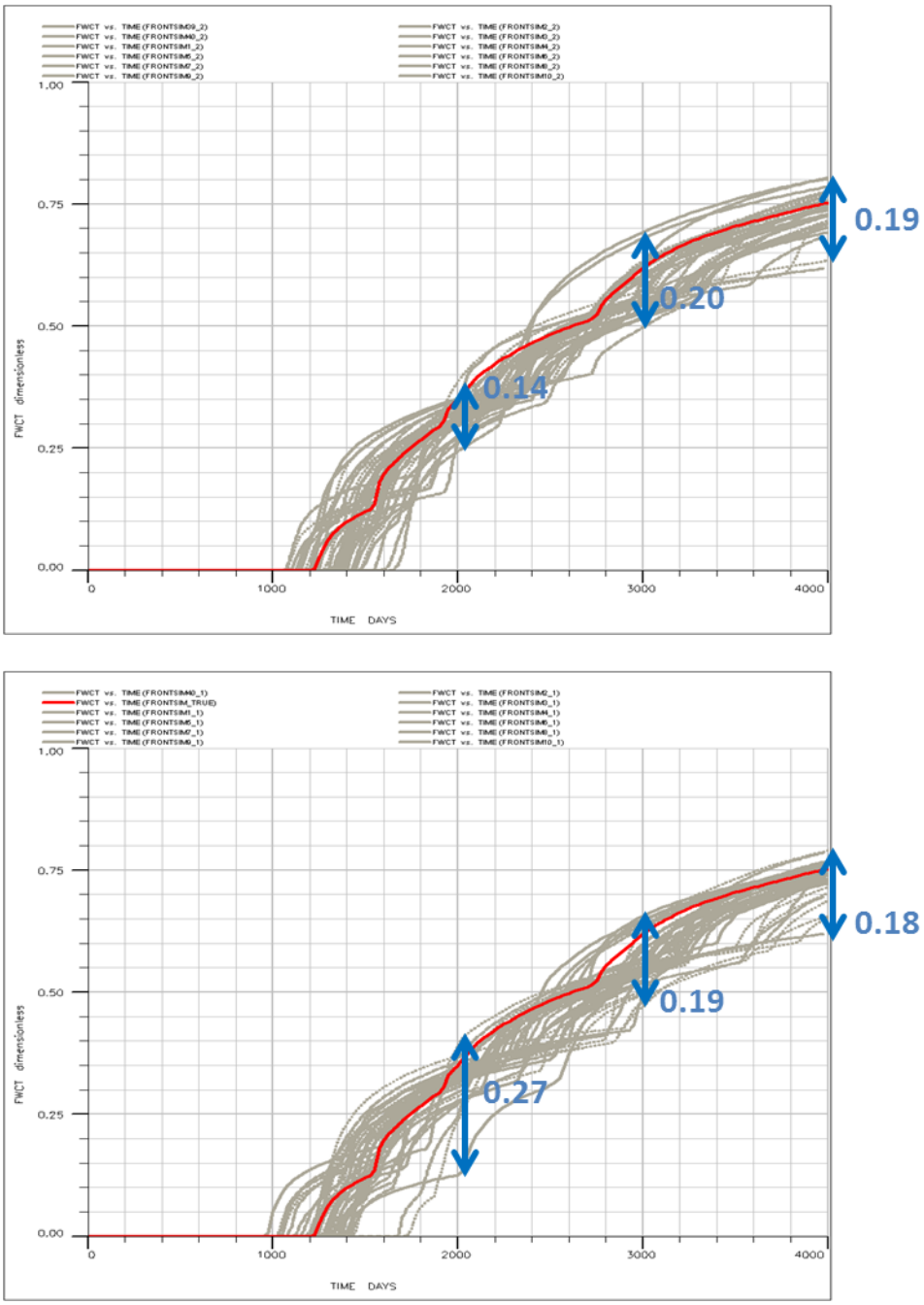


Figure 3.4. The simulated field-wide water cut responses after history-matching for (a) Plain EnKF with 40 members, (b) 40 members selected by the Multidimensional Scaling & K-means Clustering method. Red curve shows the water cut response of the reference model. The vertical blue lines dictate the width of the water cut spread. The numbers next to the lines show the numerical difference between the upper bound and lower bound water cut values at the particular time.

Table 3.1: Water cut spread at different times

Case	WWCT Spread Width at 2000 days	WWCT Spread Width at 3000 days	WWCT Spread Width at 4000 days
Before EnKF: Initial 200 Members	0.42	0.50	0.26
Before EnKF: 40 Members Selected by Spectral Clustering	0.38	0.50	0.20
Before EnKF: 40 Members Selected by MDS & K-Means Clustering	0.38	0.45	0.23
After EnKF: 40 Members Selected Randomly	0.14	0.20	0.19
After EnKF: 40 Members Selected by Spectral Clustering	0.12	**0.30**	0.20
After EnKF: 40 Members Selected by MDS & K-Means Clustering	**0.27**	0.19	0.18

In the above Table 3.1, the numbers show the numerical difference between the upper bound and lower bound water cut values at the particular time for each case. Referring to

the results before EnKF data assimilation, either key member selection method shows comparable water-cut spread to the reference case where 200 initial members are used. This suggests that using either technique can achieve a similar quality to that of using a larger pool of realizations. This motivates us to implement the key member selection as a pre-processing technique to EnKF. When referring to the results after EnKF, both key selection methods not only show comparable water cut spread to that of Plain EnKF where members are selected randomly, but they also show higher water cut spread values (boldfaced in the above table) at certain time.

Summary

In this chapter, we illustrate the motivations behind implementing the key initial member selection algorithms as pre-processing processes to EnKF. We use two algorithms as means to compute the distance/connectivity matrix, and subsequently cluster data points to obtain the key realizations with incremental swept pore volume at different time-of-flight as the input to the algorithms. The first algorithm is the spectral clustering method with specific steps proposed by Ng et al., 2002. The second algorithm is a combination of multidimensional scaling and K-means clustering. Both methods use Euclidean distance as the distance measure. Either key member selection method, as pre-processing to EnKF data assimilation, seems to achieve the benefits of using a larger ensemble size.

CHAPTER IV

CHARACTERISTICS OF ENSEMBLE KALMAN FILTER COVARIANCE LOCALIZATION SCHEMES

Creation of the Synthetic Case

This section describes the workflow of creating the synthetic case for the covariance localization study.

Table 4.1 below shows the specifications of each of the variograms used in the Sequential Gaussian Simulation (SGsims) routines employed to create the realizations for the low, medium and high heterogeneity cases respectively.

During the process of creating these realizations, a target histogram is used to condition the log permeability at the wells. The specifications of the target histogram used are outlined in the following Table 4.2.

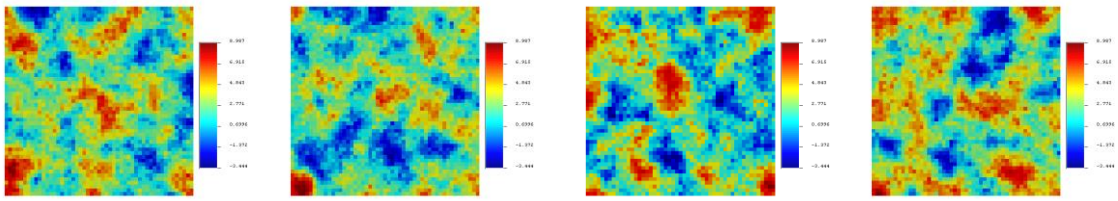
These realizations will be later used in the EnKF covariance localization study. Referring to Table 4.1 again, we should note that the maximum range for the high continuity case is 900. This value will be referred to in the later Distance-based covariance localization study.

Table 4.1: Specifications of Variograms Used in Sequential Gaussian Simulation

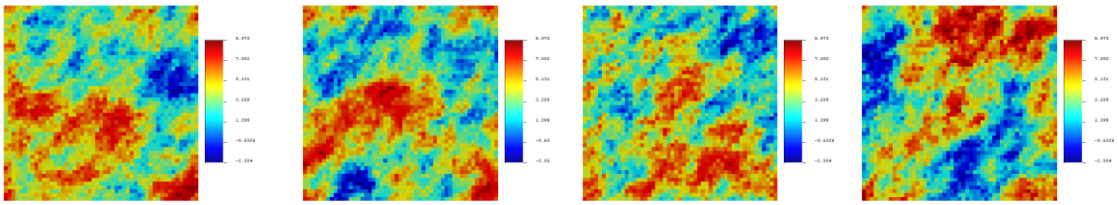
	Ranges				Angles	
Case	**Max**	**Med**	**Min**	**Anisotropy Ratio**	**Azimuth**	**Dip**
High Continuity	900	300	300	3.00	45	0
Medium Continuity	900	500	500	1.33	45	0
Low Continuity	500	500	500	1.00	0	0

Table 4.2: Target Histogram Specifications

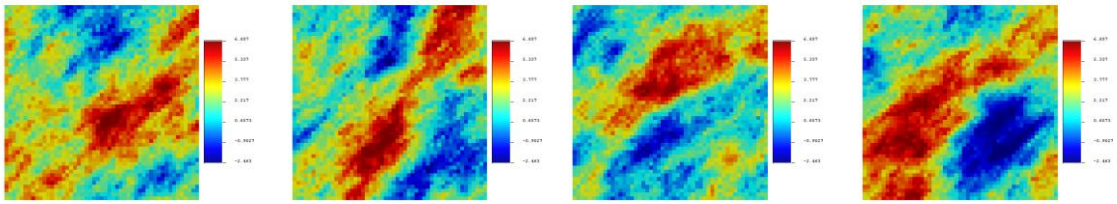
Mean	Variance	Min	Max	Bin
2.5	2.25	-3.4	8.4	0.001



(a) low heterogeneity of permeability



(b) medium heterogeneity of permeability



(c) high heterogeneity of permeability

Figure 4.1. Four initial realizations sampled from the cases of: (a) low heterogeneity of permeability, (b) medium heterogeneity of permeability, and (c) high heterogeneity of permeability.

Figure 4.1 above shows four samples of the created realizations taken from each of the low, medium and high permeability heterogeneity scenarios respectively.

The following Figure 4.2 shows the histogram of a sample realization for each of the low, medium and high heterogeneity of permeability scenarios. As mentioned earlier in Chapter II, EnKF is optimal for Gaussian models and linear model systems (Devegowda

et al., 2009). Therefore, when building the synthetic cases, one of our primary concerns is to utilize a suite of individually normally distributed members. The histograms of some of the ensemble members are examined in order to check for its normality.

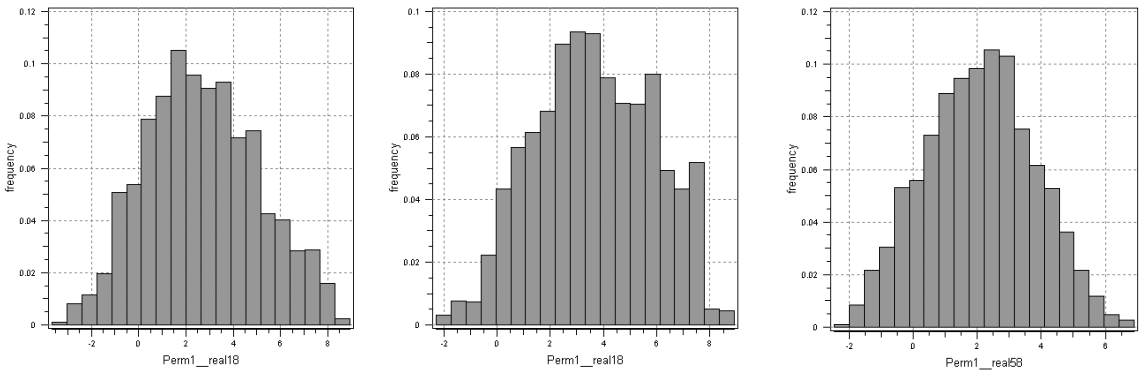


Figure 4.2. The histograms of a sample realization for each of the three cases: the low heterogeneity case (on the left), medium heterogeneity case (in the middle) and high permeability heterogeneity case (on the right) respectively. It appears that each case has a suite of individually (close to) normally distributed members.

Simulation Time of the 5-Spot Case

The preliminary synthetic field to be discussed in the following sections has the following specifications and EnKF parameters:

•High heterogeneity of permeability

•51x51x1 grid system simulating 1530 ft X 1530 ft reservoir field.

•5-Spot: 4 producers and 1 injector for water-cut matching

•Unfavorable mobility ratio with oil viscosity : 1.6 cp and water viscosity : 0.8 cp

•1 Injection Rate at 200 RESV and, 4 Production rates at 50 RESV each

•Forward simulation is from 0 day to 4000 days

•Number of ensemble members: 40

•Water-cut data assimilation at 100 day intervals from 0 day to 4000 days

•Water-cut measurement error: 8%

•State variables : (log permeability, pressure, water saturation, and water-cut)

In order to investigate the optimal simulation time for the study, a sensitivity run is conducted. A comparison of the water-cut matching results between the simulation time of 4000 days and 5400 days is shown in Figure 4.3 below. The water-cut data assimilation intervals are maintained at 100 days for both simulation schedules. Both simulation schedules of 4000 days and 5400 days have yielded comparable history matching quality.

In order to help quantity the history matching quality, the root mean square (RMS) is used as a measure to see how well overall the ensemble of computed responses are capturing the observed data. The RMS value is defined as:

$$RMS = \sqrt{\frac{1}{N_e}\sum_{i=1}^{N_e}(d_{cal,i} - d_{obs})^2} \tag{4.1}$$

where N_e is the number of ensemble members, $d_{cal,i}$ is the computed dynamic response by each of the ensemble members, d_{obs} is the observed data at the particular well, k is the assimilation time step.

The RMS is computed at each assimilation time step. The RMS values referred to in the later water-cut history matching results are computed by summing the RMS computed through the assimilation time. In general, the smaller the RMS value, the closer are the ensemble computed responses to the observed data.

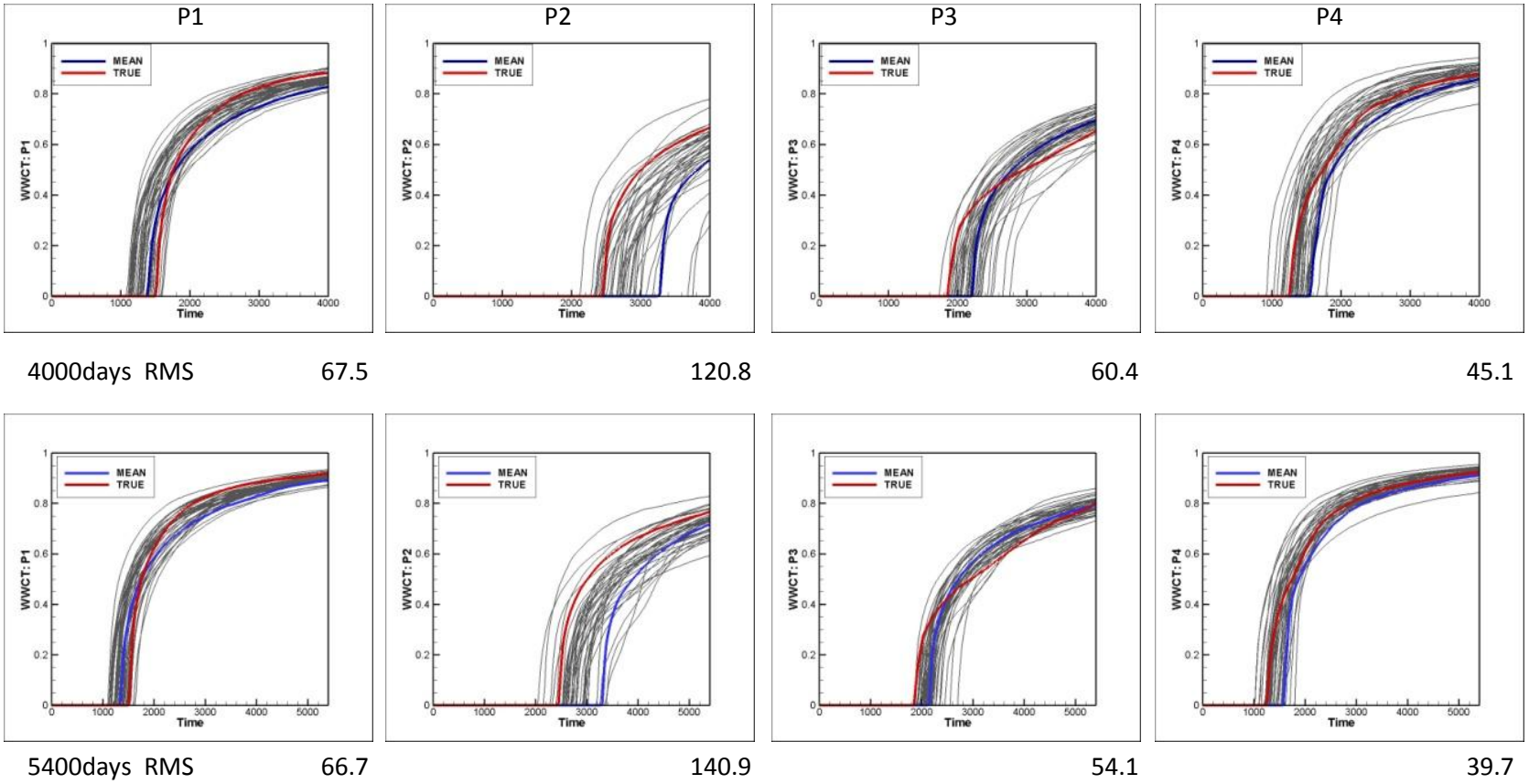
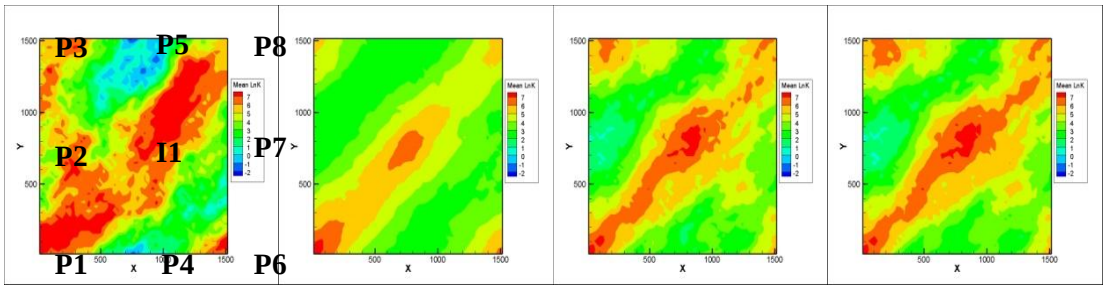


Figure 4.3. The EnKF water-cut matching results of each of the four producers for the total assimilation time of 4000days (first row) and that of 5400days (second row). The number under each plot is the root mean square (RMS) value between each ensemble response and the observed data through the entire assimilation time. Red line dictates the reference model response while blue line dictates the mean of the ensemble responses.

In addition, the final updated permeability field of the 4000-day assimilation time model has more resemblance to the true permeability field compared to the 5400-day assimilation time model as seen on Figure 4.4 below. Therefore, compared to the assimilation time of 5400 days, a shorter assimilation time of 4000days is recommended because: (a) history match quality is not sacrificed as seen in Figure 4.3 and (b) the quality of preserving the geological information is comparable to that of 5400-day model as seen in Figure 4.4 below.



(a) True Perm (b) Initial Perm Mean (c) After Perm 4000days (d) After Perm 5400days

Figure 4.4. Mean of the ensemble permeability fields; (a) Reference permeability, (b) Initial ensemble mean of permeability, (c) Mean of the final updated ensemble mean at 4000 days, (d) Mean of the final updated ensemble mean at 5400 days.

Sensitivity Study of Water-cut Measurement Error of the 5-Spot Case

The measurement error referred to in this study is a vector of the unknown measurement errors and is assumed a Gaussian distribution characterized by the measurement error variance and zero mean. This measurement error contributes to the description of the data noise, ε as in Eq. 2.4 outlined earlier in Chapter II, where

$$d_{obs,k} = d_{true} + \varepsilon_k = Hy_k + \varepsilon_k \tag{2.4}$$

The vector $d_{obs,k}$ as seen in Eq. 2.4 typically consists of measurements at the well locations, such as measurements of water-cut, gas-oil ratio and bottom-hole pressure. In this study only the well water-cut and bottom-hole pressure are concerned.

Having decided on the assimilation time for the covariance localization study, in order to investigate the sensitivity of EnKF performance to variation in measurement errors (water-cut), several EnKF runs are made at different water-cut measurement errors at 1%, 5% and 8% respectively. These percentages in measurement errors are the standard deviations of the measurement error distribution. Results are as below.

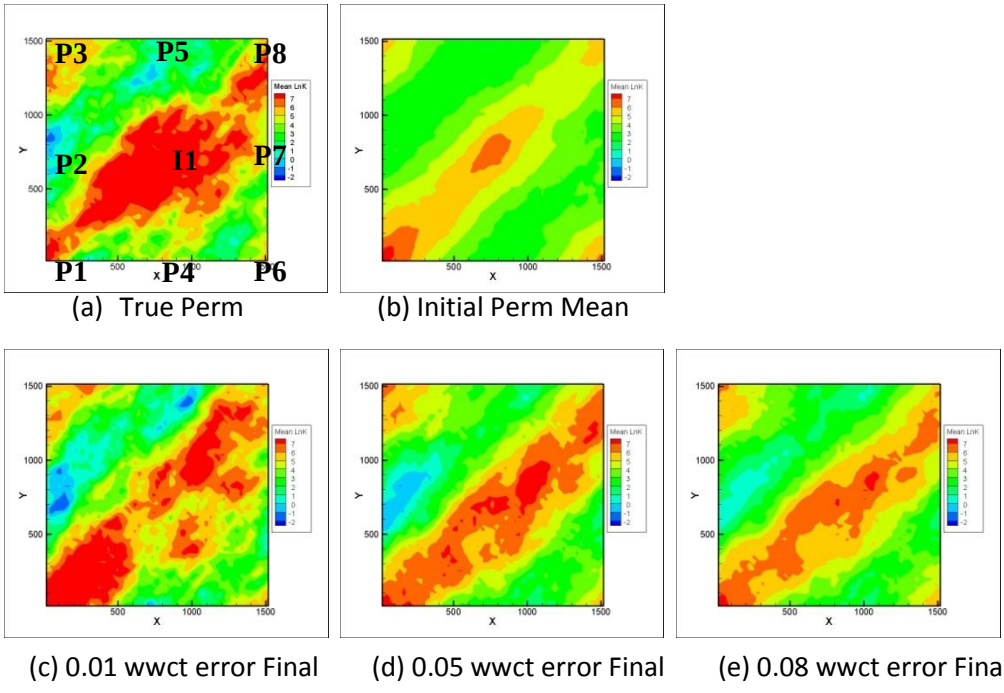


Figure 4.5. Mean of the ensemble permeability fields; (a) Reference permeability, (b) Initial ensemble mean of permeability, (c) Mean of the final updated ensemble mean for water-cut (WWCT) measurement error of 1%, (d) Mean of the final updated ensemble mean for water-cut measurement error of 5%, (e) Mean of the final updated ensemble mean for water-cut measurement error of 8%.

A comparison of the permeability fields shown in Figure 4.5 (c – e) to the true permeability field (Figure 4.5a) indicates that the geological information is not preserved adequately. Subsequently, we revise our approach and use a 9-spot pattern (with water-cut and bottom-hole pressure being assimilated) as opposed to using the 5-spot pattern (with only water-cut being assimilated).

As a result of this change in strategy, the number of measurements increases with the decrease in the distance between the observations. This is done in hope of reducing some of the spurious terms in the cross-covariance matrix.

The 9-Spot Case with Different Degrees of Anisotropy

All cases described so far are implemented on a synthetic field with high heterogeneity. In order to test the sensitivity of EnKF performance to the degree of anisotropy, we will also create two additional synthetic cases: (1) a field with low anisotropy and (2) a field with medium anisotropy. The EnKF and localization scheme will be applied to these new cases and the results will be compared to the high anisotropy case.

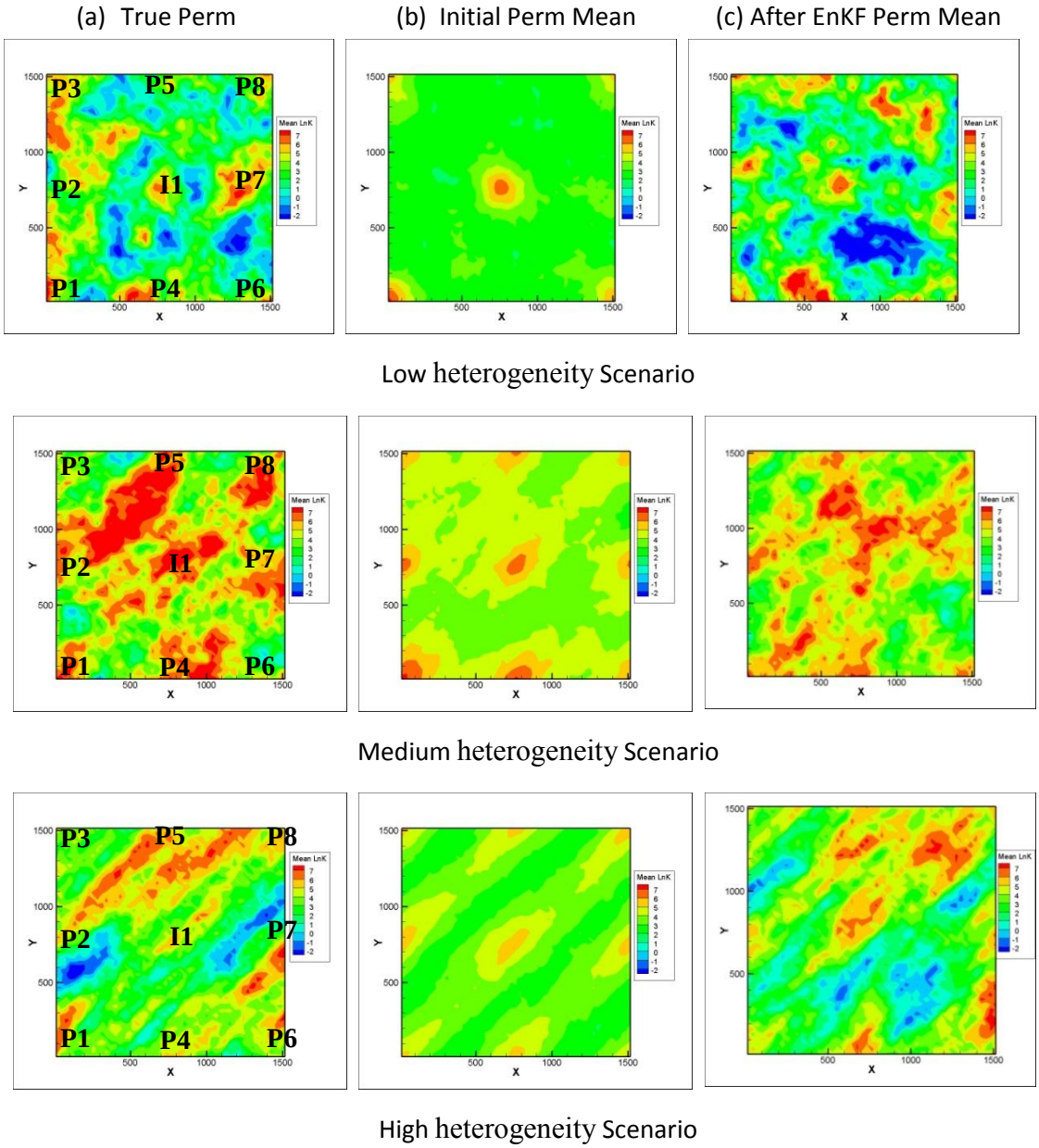


Figure 4.6. The comparison between the updated permeability fields after ENKF and the initial permeability fields at different anisotropy. It shows the mean of the ensemble permeability fields of: (a) Reference model, (b) Initial mean of the ensemble models, (c) Final updated mean of ensemble models for the low heterogeneity case (first row), medium heterogeneity case (second row), and high heterogeneity case (bottom row) respectively.

Figure 4.6 shows the reference model permeability field, the mean permeability field of the initial model and the final updated model of all three cases. Figure 4.6 (last row)

shows that the final updated permeability of the highly heterogeneous case has a more defined spatial distribution of high permeability streaks and low permeability patches. We believe that these features will be useful in distinguishing the varying performance of parameter estimation using the different covariance localization schemes later.

In the following sections, we deliberately set out to create a synthetic field with high permeability channels and low permeability barriers. This will enable us to see the streamline regions of more extreme densities. We think we will be able to showcase the advantages of using streamline based covariance localization schemes for these types of reservoir settings.

Final Revised 9-Spot Synthetic Case

Our initial synthetic model was modified based on the following;

1. Optimum simulation time is 4000 days.
2. Optimum pattern is 9-spot.
3. Optimum anisotropy ratio is high.

All these modifications resulted in a synthetic model with the following characteristics:

- Permeability field of high anisotropy
- 51x51x1 grid system simulating a 1530 ft X 1530 ft synthetic reservoir field
- 8 producers and 1 injector for water-cut and bottom-hole pressure matching
- Unfavorable mobility ratio with oil viscosity : 1.6 cp and water viscosity : 0.8 cp
- 1 Injection Rate at 200 RESV, 8 Production rates at 25 RESV each

- Number of ensembles: 40
- Both water-cut and bottom-hole pressure data assimilation at 100-day intervals from 0 day to 4000 days
- Water-cut measurement error : 5% (Best combination of measurements errors is to be decided in the next section)
- Bottom-hole pressure measurement error: 50 psi (Best combination of measurements errors is to be decided in the next section)
- Forward simulation is from 0 day to 4000 days
- State variables updated : (log permeability, pressure, water saturation, and water-cut)

Sensitivity Runs for Determining Optimal Water-cut and Bottom-hole Pressure Measurement Error Combination

In reality, the errors of measurements such as water-cut and bottom-hole pressure depend mainly on the quality of the measurement device used. Bottom-hole pressure gauges are fairly accurate. Quartz sensor type pressure down-hole gauges have an accuracy up to plus or minus 0.02 %, which means, for example, maximum error of 1psi at an operating bottom-hole pressure of 5000 psi (Spartek Systems Geophysical Instrumentation, 2008). We conduct the sensitivity of the assimilation by changing the data measurement errors. Three combinations of measurement errors of water-cut and bottom-hole pressure at: (a) 5% and 5 psi, (b) 5% and 1psi, and (c) 8% and 150 psi are examined. All cases have both water-cut and bottom-hole pressure matching results examined. The water-cut history matching is first discussed as the following.

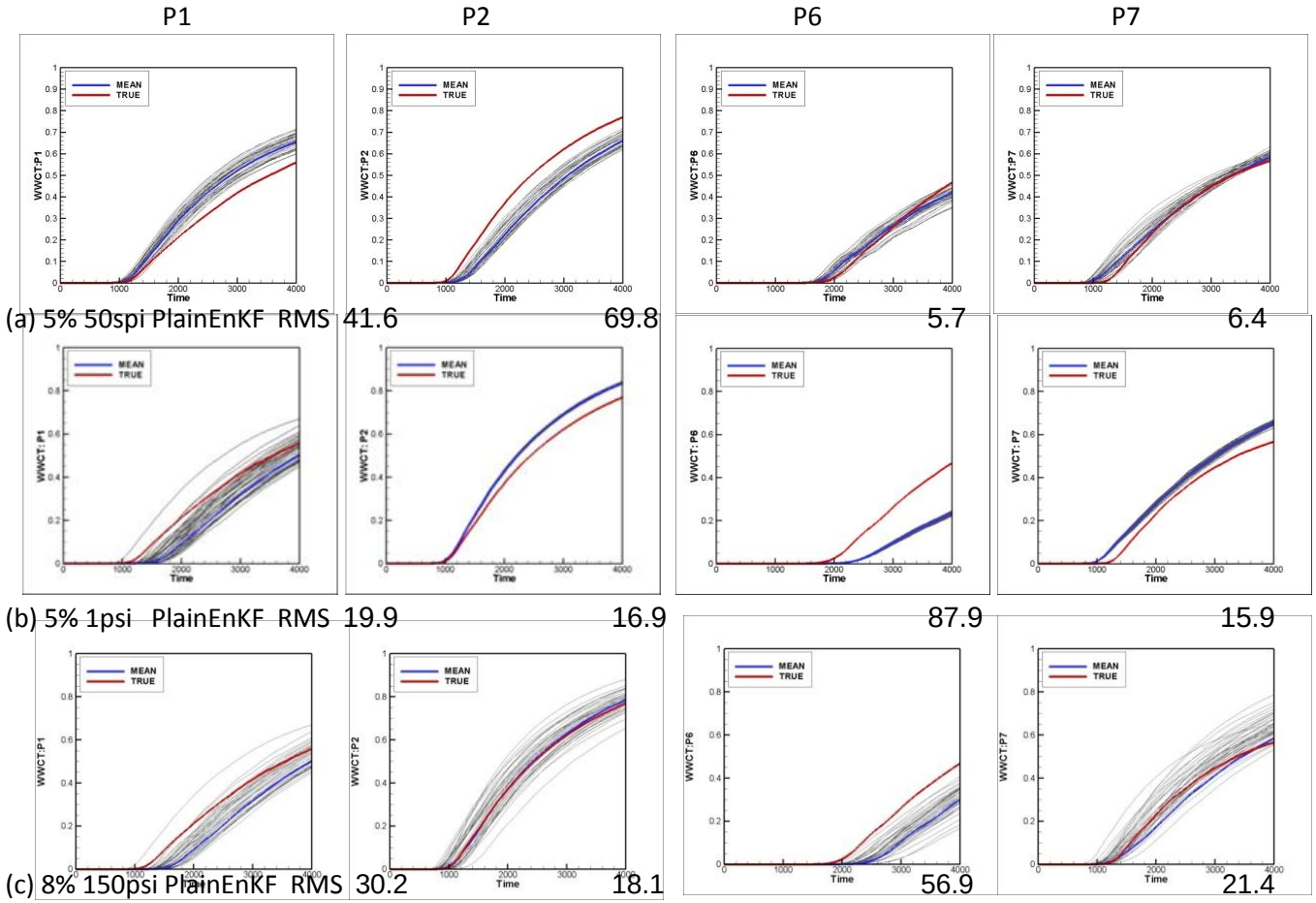


Figure 4.7. The EnKF water-cut matching results of four producers: P1, P2, P6 and P7 for the cases with water-cut and bottom-hole pressure measurements errors respectively at (a) 5% and 50psi, (b) 5% and 1psi, and (b) 8% and 150psi. Red line dictates the reference model response while blue line dictates the mean of the ensemble responses.

In Figure 4.7, the number below each plot is the root mean square (RMS) value as defined earlier. The combinations of water-cut and bottom-hole pressure measurement errors at 8% and 150psi [case (c)] respectively show overall satisfactory results, which is, low RMS values and a closer mean of ensemble responses to the reference model response. However, as mentioned earlier, pressure down-hole gauges are fairly accurate, so a bottom-hole measurement error of up to 150 psi [case(b)] is probably an over-kill.

The model responses converge for case (b) with 5% and 1psi measurement errors. This phenomenon is known as the ensemble collapse.

It seems that when less variance of the measurement errors is assigned, the final updated ensemble responses become more affected by the observations in terms of the model spread converging towards the observation data.

Therefore, it is necessary to assign the appropriate measurement error variance to prevent ensemble collapse and at the same time maintain model uncertainty. This is the motivation for experimenting with different cases to obtain the optimal combination of measurement (water-cut and bottom-hole pressure) errors to use for the later EnKF study.

The history matching results of the case with measurement errors at 5% (water-cut) and 50 psi (bottom-hole pressure) [case (a)] and that at 8% (water-cut) and 150 psi (bottom-hole pressure) [case (c)] are comparable. Since the measurement errors at 5% (water-cut) and 50 psi (bottom-hole pressure) are more realistic with respect to typical field scenarios, this combination is opted for the later study. In the following section, bottom-hole history matching results for the three cases will be discussed.

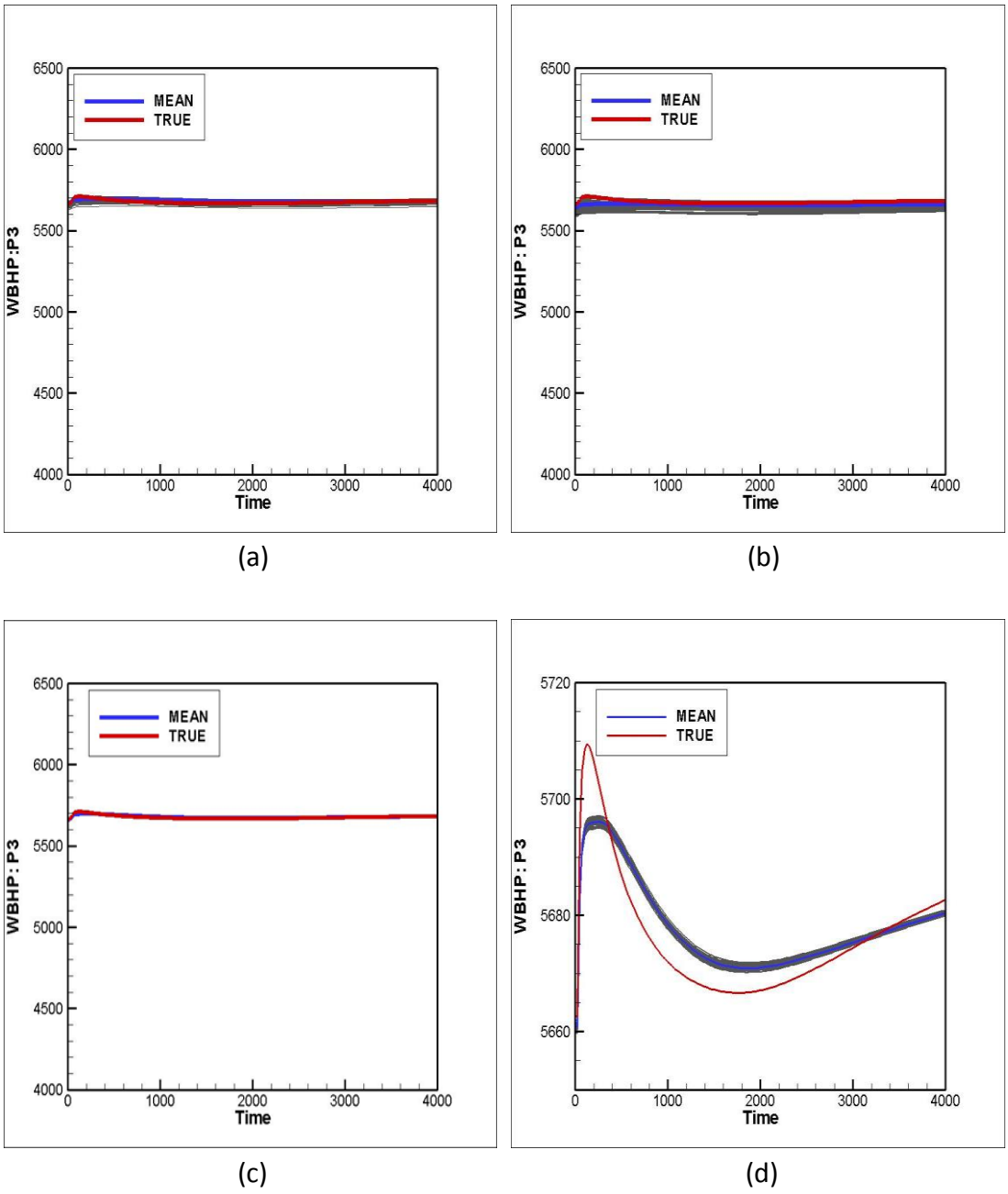


Figure 4.8. The EnKF bottom-hole pressure matching results of producer: P3 for the cases with (a) water-cut error at 5% and bottom-hole pressure error at 50psi, (b) water-cut error at 8% and bottom-hole pressure error at 150psi, (c) water-cut error at 5% and bottom-hole pressure error at 1psi, and (d) the refined plot of (c) showing the collapse of the ensemble responses for the case with water-cut error at 5% and bottom-hole pressure error at 1psi . Red line dictates reference model response while blue line dictates mean of the ensemble responses.

Figure 4.8 shows that the bottom-hole pressure matching results for well P3 for all three cases. All cases have comparable bottom-hole pressure profile matches.

Ensemble Kalman Filter (EnKF) is implemented in the final revised 9-Spot synthetic field with the following schemes: (i) no covariance localization (Plain EnKF), (ii) distance based covariance localization (EnKF-DS), (iii) streamline trajectory based covariance localization (EnKF-ST), (iv)streamline sensitivity based covariance localization (EnKF-SS), and (v) hierarchical EnKF (EnKF-HR).

The characteristics and methodology of each localization scheme is outlined in the following section. The history matching results of implementing each covariance localization scheme is also attached.

Motivation and Methodology for Covariance Localization Study

As outlined earlier in Chapter II, the equation used to update the state vector during EnKF is:

$$\mathbf{y}_k^u = \mathbf{y}_k^P + C_k^P H_k^T (H_k C_k^P H_k^T + C_{D,k})^{-1} \left(\mathbf{d}_{obs,k} - \mathbf{H}_k \mathbf{y}_k^P \right) \tag{2.16}$$

while the Kalman Gain is:

$$K_k = C_k^P H_k^T (H_k C_k^P H_k^T + C_{D,k})^{-1} \tag{2.17}$$

In the context of covariance localization, $\boldsymbol{\rho} \circ$, as known as either the localization function or multiplier function in this thesis, is introduced to condition the Kalman Gain. The symbol '$\circ$' is an element-by-element multiplication operator known as the Schur product

(Gaspari and Cohn, 1996). EnKF update with the implementation of covariance localization is revised as:

$$\mathbf{y}_k^u = \mathbf{y}_k^P + \rho \circ C_k^P H_k^T (H_k C_k^P H_k^T + C_{D,k})^{-1}\left(\mathbf{d}_{obs,k} - \mathbf{H}_k \mathbf{y}_k^P\right) \tag{4.1}$$

Each localization scheme has its own methodology of arriving at the multiplier, with the common theme to localize the cross covariance matrix.

As outlined earlier in Chapter II, the cross covariance matrix is defined as:

$$\mathbf{C}^P H^T = \begin{pmatrix} \mathbf{C}_{\mathbf{m}^{sta}} & \mathbf{C}_{\mathbf{m}^{sta},\mathbf{m}^{dym}} & \mathbf{C}_{\mathbf{m}^{sta},\mathbf{d}^{cal}} \\ \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} & \mathbf{C}_{\mathbf{m}^{dym}} & \mathbf{C}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} \\ \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{sta},\mathbf{d}^{cal}} & \mathbf{C}^{\mathbf{T}}_{\mathbf{m}^{dym},\mathbf{d}^{cal}} & \mathbf{C}_{\mathbf{d}^{cal}} \end{pmatrix}^p \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} C_{m^{sta},d^{cal}} \\ C_{m^{dym},d^{cal}} \\ C_{d^{cal}} \end{pmatrix}^P \tag{2.12}$$

Within the cross-covariance matrix, it is the sub-matrix $C_{m^{sts},d^{cal}}$ that is the main driver for the updates in the static variables (parameters). This sub-matrix captures the cross-covariance between the static variables and observation data. As the parameters and observation data concerned in this thesis are log permeability, and water-cut along with bottom-hole pressure respectively, the sub-matrix $C_{m^{sts},d^{cal}}$ can be outlined as the following:

$$C_{m^{sts},d^{cal}} = \begin{matrix} \mathrm{cov}[\log(k_1), WBHP] & \mathrm{cov}[\log(k_1), WWCT] \\ \vdots & \vdots \\ \mathrm{cov}[\log(k_{N_{grid}}), WBHP] & \mathrm{cov}[\log(k_{N_{grid}}), WBHP] \end{matrix} \tag{4.2}$$

The covariance localization techniques to be discussed in the following chapters revolve around conditioning the above cross-covariance matrix. The localizing function used to

condition the cross-covariance is computed at each data assimilation time. All in all, the fundamental goal of the covariance localization schemes is to eliminate erroneous terms in the cross-covariance matrix, in an attempt to preserve the underlying geology and allow for more relevant parameter updates.

Distance Based Covariance Localization

Since permeability is a varying parameter throughout the reservoir field which is typically heterogeneous, the pressure drop is also varying among the grid blocks. As a result, the fluid production rate (i.e. oil, water production rate) and the permeability follow a non-linear relationship especially for grid blocks farther apart. Therefore, it is imperative to understand the effects of the distance between the well of observations and the grid block concerned on estimating the permeability.

The computed cross-covariance between the permeability and fluid production rate (i.e. water cut) can be overestimated due to their highly non-linear relationship especially when a smaller ensemble size is used. This overestimate of value of the cross-covariance matrix is known as the overshooting problem (Hamill et al., 2001).

As mentioned earlier in Chapter II, another substantial problem of using EnKF is caused by using small-sized ensemble members to estimate the background error covariance (Hamill et al., 2001).When background errors are underestimated, the observation is comparatively ignored and the posterior distribution can excessively resemble the prior. Hamill et al. (2001) noted that the EnKF analysis could be improved by excluding observations at great distances from the grid point of analysis.

It is therefore sensible to "cut-off" cross covariance terms at grid blocks at certain distance away from the well of observation in an attempt to remove the spurious covariance terms. One of the ways to achieve this is by analyzing the correlation of the distance between the well with the observation data and the grid block concerned.

In the following section, we will review the distance-based covariance localization as a means to improve EnKF performance by modifying the cross-covariance calculations. The mathematical formulations proposed that are used to capture this correlation will also be a point of discussion.

In order to define the correlation function *S* used in this type of covariance localization, we use a fifth order function outlined by Gaspari and Cohn (1996), which is similar to a Gaussian function where the correlations translate to the different multiplier terms across all grid blocks corresponding to each well.

The set of equations used to define the correlations are as follows:

$$\begin{aligned}
&\rho(a,b) = -\frac{1}{4}\left(\frac{b}{a}\right)^5 + \frac{1}{2}\left(\frac{b}{a}\right)^4 + \frac{5}{8}\left(\frac{b}{a}\right)^3 - \frac{5}{3}\left(\frac{b}{a}\right)^2 \text{.....}, 0 \le b \le a; \\
&\rho(a,b) = \frac{1}{12}\left(\frac{b}{a}\right)^5 - \frac{1}{2}\left(\frac{b}{a}\right)^4 + \frac{5}{8}\left(\frac{b}{a}\right)^3 + \frac{5}{3}\left(\frac{b}{a}\right)^2 - 5\left(\frac{b}{a}\right) + 4 - \frac{2}{3}\left(\frac{b}{a}\right)^{-1}, a < b \le 2a; \\
&\rho(a,b) = 0\text{.....}, b > 2a.
\end{aligned} \tag{4.3}$$

b is the distance between the well location (observation point) and the grid point underlying the model parameters, where *a* is the multiplication of the factor of $\sqrt{10/3}$ and the length scale l_c:

$$a = \sqrt{10/3} l_c \tag{4.4}$$

or $F_c = \sqrt{10/3} l_c$, since $a = F_c$ (4.5)

A correlation matrix S is defined for every grid point (i,j) in the domain:

$$S(i, j) = \Omega(F_c, \|D_{ij}\|) \quad (4.6)$$

D_{ij} denotes the Euclidean distance between grid point (i,j) and the location of observation (Hamill, Whitaker et al. 2001). It is intuitive from the above equations that the multiplier decreases from 1's at the observation locations to 0's asymptotically at grid blocks of certain distance away from the wells until the certain cut-off distance is reached. Therefore, the localizing function is basically a matrix of 0's and 1's. Fundamentally, the closer the grid point underlying the model parameters is to the well location, the higher the values of the multiplier matrix terms. This creates an effect of weighing and restricting the statistical information of the observation data to grid cells closer to the measurement location.

Distance Based Covariance Localization Data Assimilation Results

In this study, the distance based covariance localization schemes are implemented based on the assumption that the model parameters and observation data are associated within a certain cut-off radius.

The cut-off radius is chosen arbitrarily. Within this cut-off radius, a weighting function as outlined earlier proposed by Gaspari and Cohn, is implemented in a way that the terms closer to the observations receive higher weights and progressively less at farther grid blocks. Beyond that cut-off radius the weights are assigned to be zero. Subsequently, the cross-covariance matrix is associated with a localizing function.

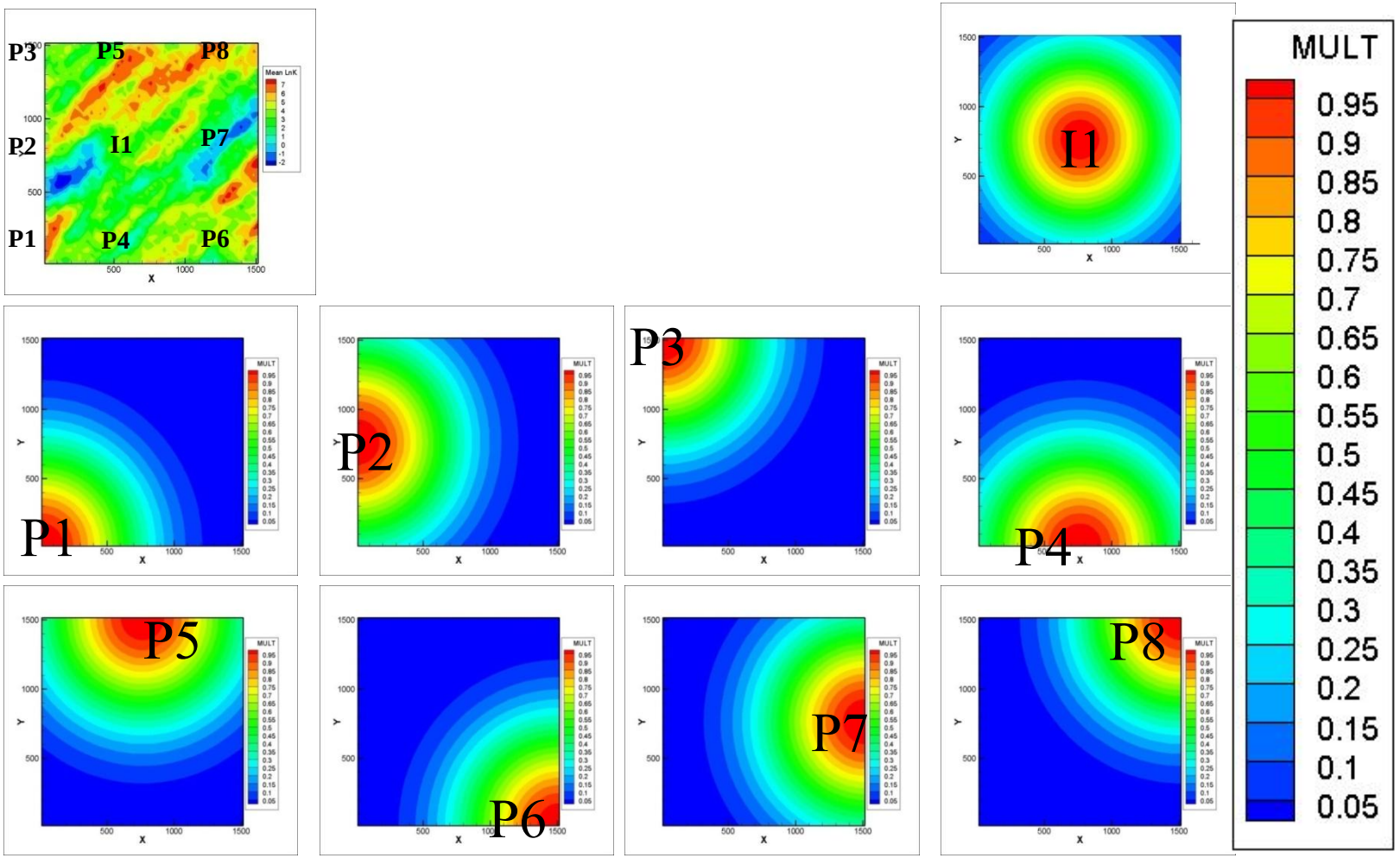
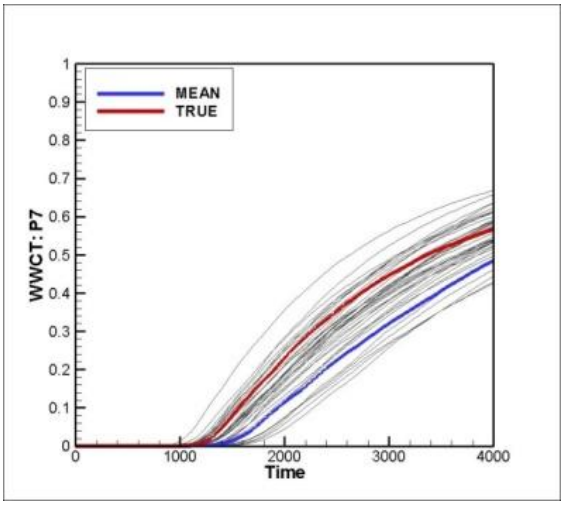


Figure 4.9. The map of the multiplier for each well out of the 8 producers and 1 injector for distance based localization at t=4000days; the top left corner is the log permeability field of the reference model.

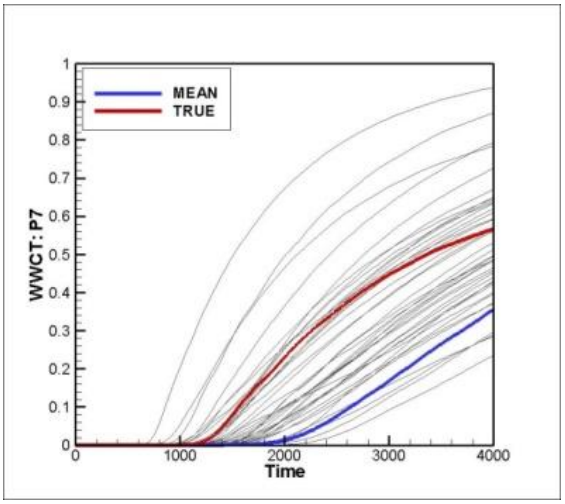
The results of implementing this type of covariance localization scheme will be discussed in the following session.

Figure 4.9 shows the spatial distribution of the multiplier values for each well at the last assimilation time *t=4000days*. These multiplier values concerned demonstrate the correlation between the model parameters (log permeability) and the observation data (water-cut). Each number on each picture depicts the specific well at which this correlation is being computed for.

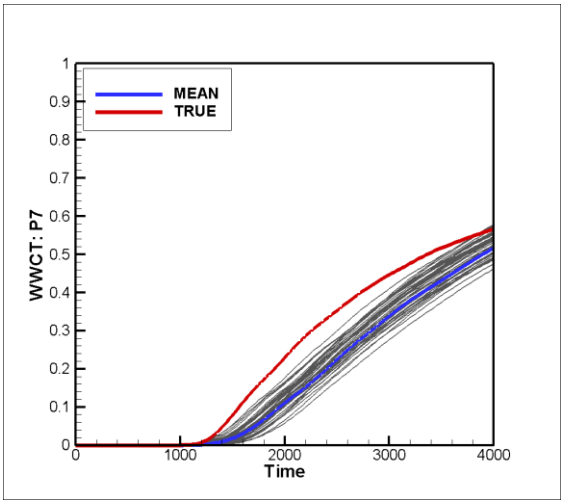
In this study, the preliminary pick for the cut-off radius is 900, in comparison to the synthetic model with dimensions of 1530 ft x 1530 ft. In addition, this preliminary guess of the cut-off radius is equal to the range (at 900) of the variogram originally used to create this reservoir model. The study is repeated with cut-off radii at 500 and 3000 with history matching results as shown next.



(a) Cut-off limit=900 RMS 22.5



(c) Cut-off limit=500 RMS 116.9



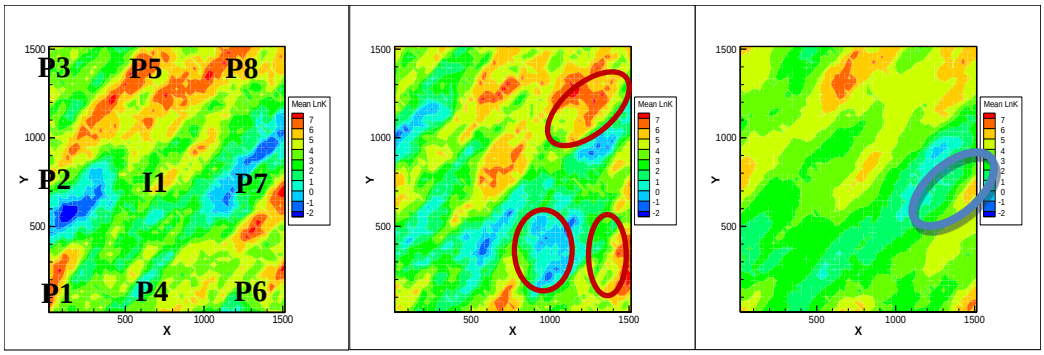
(d) Cut-off limit=3000 RMS 37.4

Figure 4.10. The EnKF water-cut matching results of a selected producer: P7. The water-cut matching results for EnKF with distance based covariance localization with cut-off limit at: (a) 900 (same as the length scale), (c) 500 (shorter than the length scale), and (d) 3000 (longer than the length scale). The number under each plot is the root mean square (RMS) value. Red line dictates reference model response while blue line dictates mean of the ensemble responses.

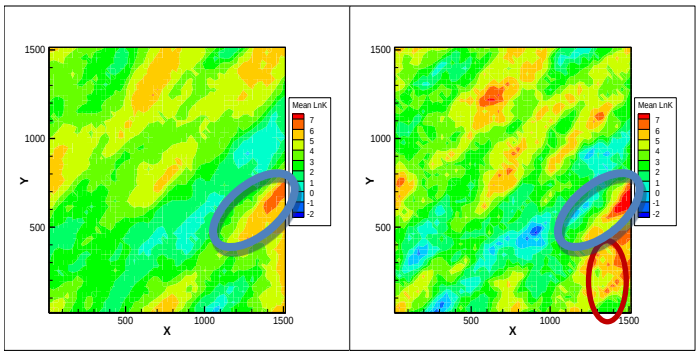
Figure 4.10 shows that the water-cut matching results for a specific well P7 among the 8 producers. From an RMS minimization point of view, among the three cases, the history match quality is best when the cut-off radius is comparable to the range (at 900) of the variogram. This case has the greatest variance of the model responses, which at certain degree reflects its ability to capture the model uncertainty. However, when the cut-off radius is much lower (at 500) or higher (at 3000), the RMS values are higher. This varying degree in performance reflects the need to choose the optimal length-scale to use.

In summary, the results show benefits of using distance covariance localization as long as the appropriate length scale is used.

However, the selection of a characteristic length scale can be aided by initial geological knowledge but overall requires trial-and-error. Furthermore, the type of covariance localization may not be consistent with the underlying heterogeneity that can include high permeability channels with long-range contributions (Arroyo et al. 2006). In order to examine if it is the case, the final updated permeability field of the above cases is compared to the reference permeability field. Results are as shown in the following figure.



(a) Reference Model (b) Plain EnKF (c) Cut-off limit = 500



(d) Cut-off limit = 900 (e) Cut-off limit = 3000

Figure 4.11. The final updated permeability for these cases: (a) the reference model, (b) the final updated permeability for the EnKF case without any covariance localization scheme; and the final updated permeability field for the EnKF case with distance based covariance localization with cut-off limit at: (c) 900 (same as the length scale), (d) 500 (shorter than the length scale), and (e) 3000 (longer than the length scale) respectively.

It is apparent from the above Figure 4.11 that there are permeability overshooting and undershooting issues (in regions circled as red) for case (b) the Plain EnKF case. However, when the Distance-based covariance localization with cut-off limit at 900 or 500 is used, the parameter overshooting and undershooting is mitigated as seen in cases (c) and (d).

In addition, the Distance-based covariance localization with cut-off limit at 500 [case(c)] has not retained the high permeability streak (circled as blue) where that at 900 [case(d)]

or at 3000 [case(e)] has. However, the case with cut-off limit of 3000 [case(e)] has a permeability overshooting (highlighted as red). Results show that the cut-off limit of 900 is the optimum cut-off limit to use among the three choices.

Hamill and Whitaker (2001) observed that the length scale chosen to achieve covariance localization is a function of the ensemble size. When ensemble replicates are small in number, the appropriate cut-off radius is expected to be smaller because short-range correlations are significantly corrupted by noise. As ensemble size increases, however, the effect of noise on the distant observations is diminished and calls for a larger radius of influence.

In order to examine the quality of the distance-based covariance localization scheme in more dimensions, the RMS values of the updated permeability with respect to the reference model permeability through the assimilation time are examined.

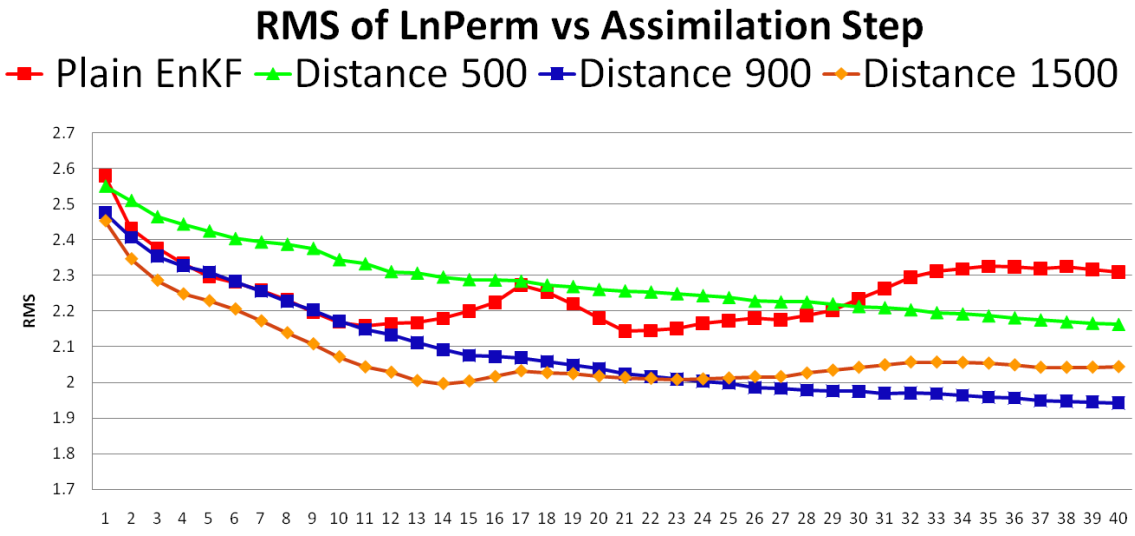


Figure 4.12. The RMS value of the final updated mean permeability with respect to the reference model permeability through time.

Figure 4.12 shows that at the final time step, the base case with the cut-off radius at 900 has the lowest RMS value especially at the late half of the assimilation time.

From the results obtained so far, the limitations of distance-based covariance localization can be summarized as the following:

i) Overall, the distance based covariance localization case mitigates the parameter overshooting and undershooting experienced in the Plain EnKF case. It is achieved only when the optimal cut-off limit is used, however.

ii) Selection of the length scale is subjective and requires trial-and-error.

Streamline Trajectory Based Covariance Localization

One of the key concerns during history matching is to change parameters that will have the most significant effects on the simulation. Therefore it is important to identify the zones where parameter changes will have the greatest impact or where the parameters and observations are mostly correlated. We can then perform the required changes in these zones preferentially to obtain more accurately updated reservoir models.

Prior information of the governing physical phenomena can be used to infer these zones. For instance in primary depletion, bottom-hole pressures are dominantly affected by reservoir parameters within the zone defined by the radius of investigation.

During water flooding, water-cut is mainly affected by the model parameters within the water swept regions, the covariance matrix can help dictate regions where the correlations

between the parameters and observations are stronger, provided that the covariance is well conditioned. The covariance matrix term $\mathbf{C}_{m^{sts},d^{cal}}$ outlined earlier is restated in the following:

$$C_{m^{sts},d^{cal}} = \begin{matrix} \text{cov}[\log(k_1), WBHP] & \text{cov}[\log(k_1), WWCT] \\ \vdots & \vdots \\ \text{cov}[\log(k_{N_{grid}}), WBHP] & \text{cov}[\log(k_{N_{grid}}), WBHP] \end{matrix} \tag{4.2}$$

However, when the covariance matrix is not very well conditioned, other physical phenomena can be utilized to identify these preferential zones. Therefore, it is sensible to investigate methods to condition the EnKF covariance matrix.

Using streamlines the water swept zones can be identified at different time-of-flight values. The EnKF covariance conditioned by the streamline information can be used to deduce zones where the correlation between the production data and the parameters is stronger (Emanuel and Milliken 1998). All in all, the primary focus of streamlines and the information derived therein is to make adjustments aimed at these zones.

In the context of streamline trajectory localization, in order to identify regions corresponding to each observation, it is necessary to identify the grid blocks intersected by a streamline originated from the observation location.

The result is that only the terms related to the selected grid blocks will be retained within the column of this covariance matrix corresponding to this particular observation (Arroyo et al. 2006). The rest of the terms within the matrix will be set to zero. In turn, the area of influence for all members is defined by stacking the selected grid cells of each ensemble. Subsequently the matrix terms related to the selected grid blocks are used to condition the

Kalman Gain. Results of implementing this kind of covariance localization on EnKF history matching on a 9-Spot synthetic field case are discussed in the following section.

Streamline Trajectory Based Covariance Localization Data Assimilation Results

The covariance localizing function for each well is plotted in the following figure at *t=4000days*, the last data assimilation time step. As seen in Figure 4.13 below, the red regions indicate the multiplier values of 1's while the blue regions dictate 0's. Fundamentally, only the grid blocks intersected by the streamlines will have the covariance matrix components retained.

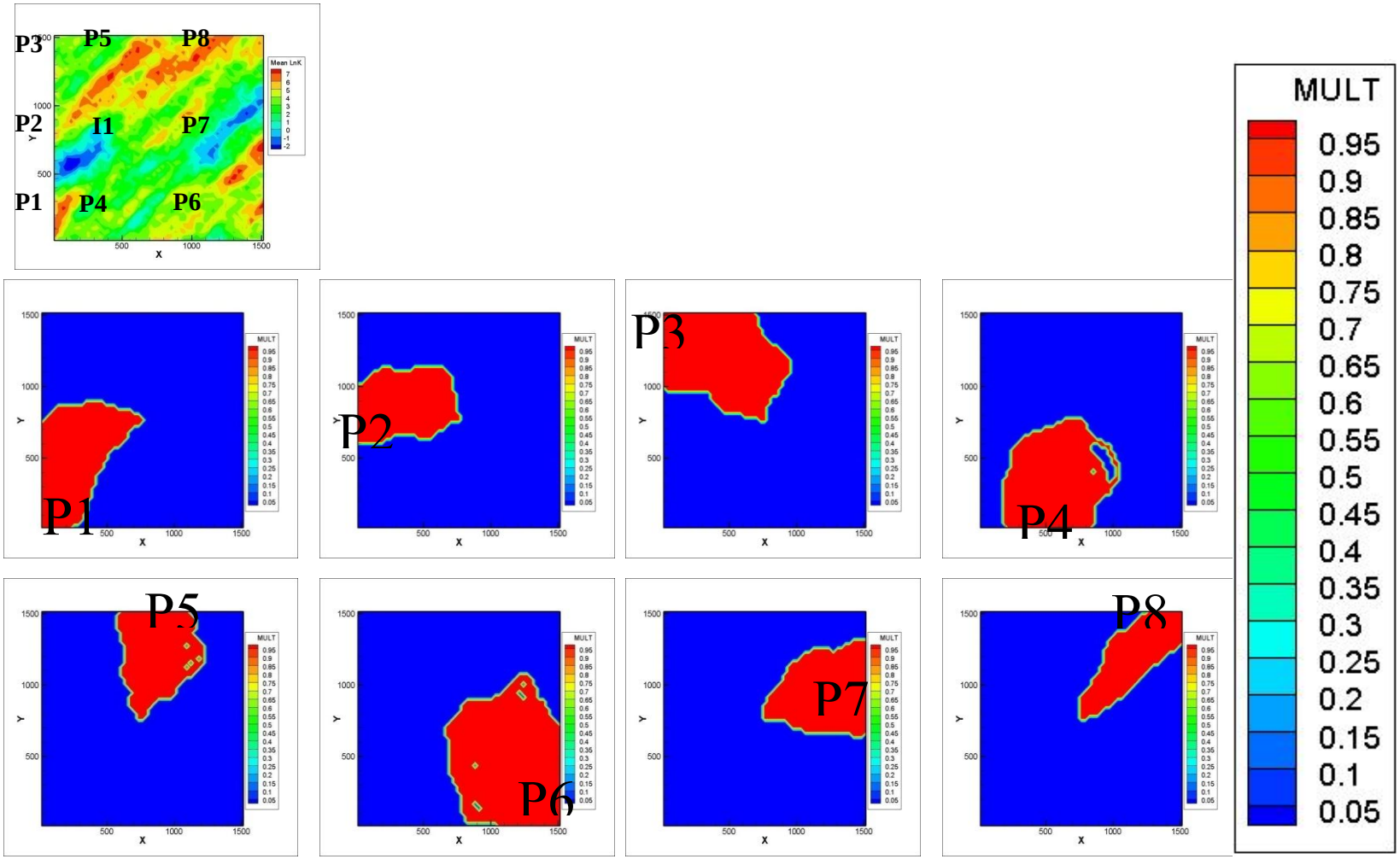


Figure 4.13. The map of the multiplier values of each observation at the producers for streamline trajectory based localization at the final assimilation time step; the top left corner is the permeability field of the reference model.

Streamline Sensitivity Based Covariance Localization

The following section describes the application of streamline-derived sensitivities to covariance localization in EnKF implementation. Using a test case, we demonstrate the application of this technique to a synthetic case of high heterogeneity. In summary, sensitivities relate changes in the production response to variations in the parameters.

In the context of petroleum reservoir inverse problems, customarily we need to determine the resulting change of data when the model is slightly changed. This is referred to as the sensitivity of the data to the model. The sensitivity values typically capture the sensitivities of changes in dynamic variables (i.e. pressure, saturation, etc) to small perturbations in model parameters (i.e. permeability). The relationship between these changes has to be derived from the partial differential equations used for the flow and transport problem.

Parameter sensitivities are often used in data integration using inverse modeling algorithms to estimate the spatial distribution of the unknown parameters e.g. porosity and permeability. Sensitivities capture the relationship between the changes in the production data and the changes in the parameters namely permeability in this synthetic field case.

In this study, the localization function ρ for streamline sensitivity localization is defined as:

$$\rho_{i,j} = \left| \frac{s_{i,j}}{\max_j(s_{i,j})} \right| \tag{4.7}$$

where *i* refers to the grid point and j refers to the specific observation data.

All in all, the sensitivity values refer to the sensitivity of production data (water-cut) with respect to the variation in model parameter (log permeability). However, the bottom-hole pressure sensitivities are not integrated to capture the correlation between bottom-hole pressure and the parameters in this study. Instead, only the streamline trajectory values (either 0's or 1's) are used to capture the correlation between bottom-hole pressure and log permeability.

Results of implementing this kind of covariance localization on EnKF history matching on the 9-Spot synthetic field case are discussed in the following section.

Streamline Sensitivity Based Covariance Localization Assimilation Results

The cross-covariance multiplier for each observation is plotted in the following figure at *t*=4000days. It shows the varying strength of the correlation between model variables and the water-cut data within the streamline region.

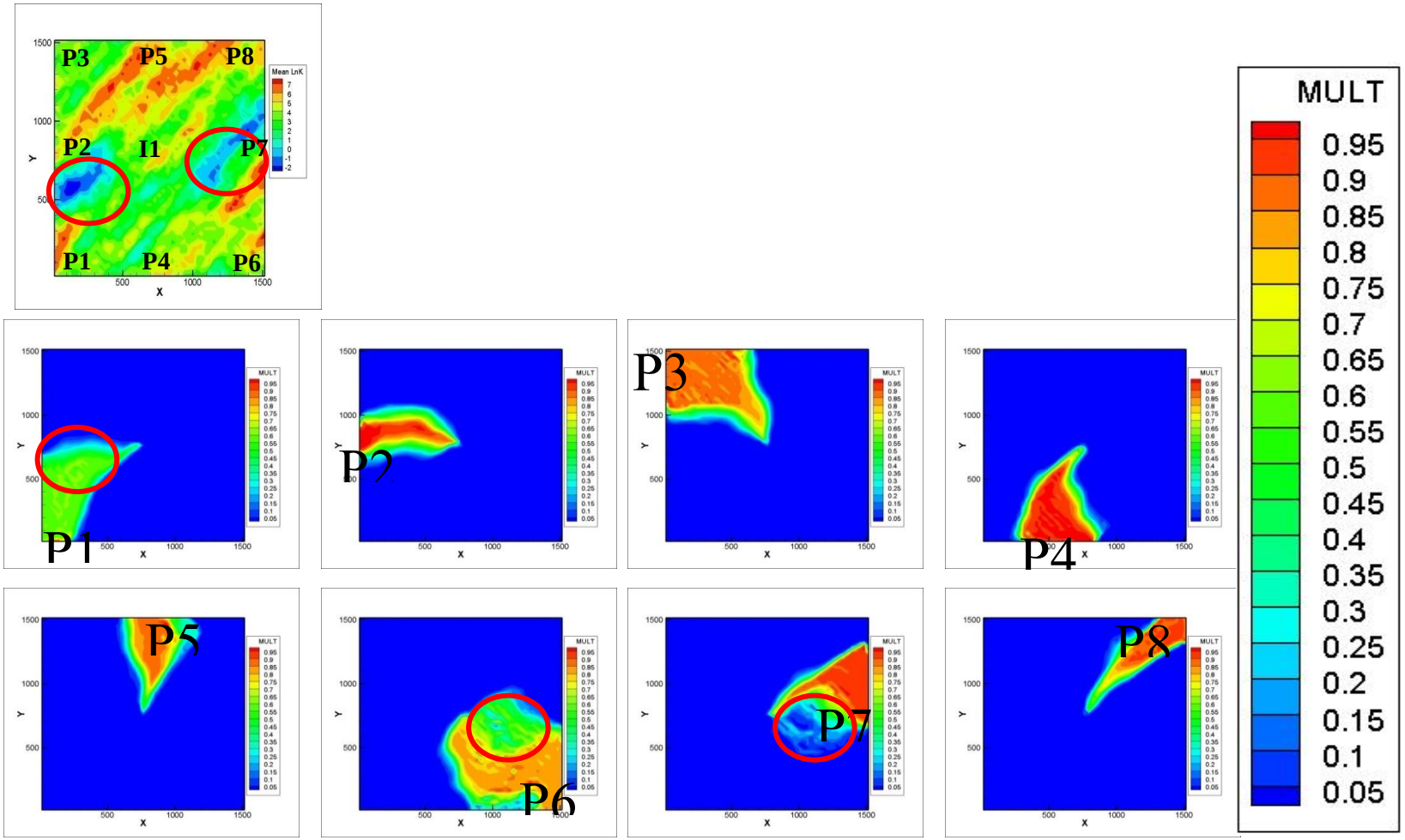


Figure 4.14. The map of the multiplier values for each producer after history matching using EnKF with streamline sensitivity based covariance localization at *t*=4000days; the top left corner is the log permeability field of the reference model

Figure 4.14 shows the distribution of the localizing function values used to condition the cross covariance between water-cut and log permeability for each producer at *t=4000days*. The circled regions in P1, P6 and P7 indicate smaller multiplier values (less correlation between water cut and log permeability). This is in agreement with the lows located at the circled regions on the reference model of log permeability. This is one additional feature of using streamline sensitivity covariance localization when compared to streamline trajectory covariance localization in that it is able to capture the relative influence within the streamline zones.

As mentioned in the earlier section,

$$\rho_{i,j} = \left| \frac{s_{i,j}}{\max_j(s_{i,j})} \right| \tag{4.8}$$

where *i* refers to the grid point and j refers to the specific observation data. $S_{i,j}$ is obtained by summing the sensitivities for all members of each producer normalized by the maximum sensitivity value for that specific observation data. The multiplier ρ is computed at each update step to reflect the changes in this kind of field conditions through time.

In order to take the relative correlation into account, the sensitivity of fractional flow to model parameters (i.e. permeability) is used. To this end, sensitivities to the model parameters are computed along the streamlines during the forward simulation.

These streamline-derived sensitivities (at all grid blocks) of all ensemble realizations are used to quantify the localization function ρ. The streamline-sensitivity based covariance localization defers from that of trajectory based in that: the varying degree of correlation

between the model parameters the model variables and the calculated responses is captured.

Hierarchical Ensemble Kalman Filter

Hierarchical Ensemble Kalman Filter (Hierarchical EnKF) involves splitting the initial members into groups of members. According to Anderson 2004, the procedures of implementing Hierarchical EnKF are outlined in the following:

Assume we have m groups of n-member ensembles (total 'm x n' members). We use linear regression to compute the increment in a state variable x (e.g. permeability) given increments for an observation variable y (e.g. water rate), where m sample values of the regression coefficient, β , are available. The regression coefficient is computed by:

$$\beta_i = \sigma_{x,y} / \sigma_{y,y} \tag{4.9}$$

$\sigma_{x,y}$ is the prior sample covariance of the state variable *x* and the observation variable *y*.

$\sigma_{y,y}$ is the prior sample variance of two observed variables with each computed using the n members from each of the different m groups.

The implementation of Hierarchical EnKF has the following procedures (Anderson 2004):

i) Divide the initial realizations into m groups of n-member ensembles (total 'm x n' members). The Hierarchical EnKF algorithm attempts to minimize the multiplier ρ using the following equation:

$$\rho_{\min} = \sqrt{\sum_{j=1}^{m} \sum_{i=1, i \neq j}^{m} (\alpha\beta_i - \beta_j)^2} \tag{4.10}$$

β is the regression factor from the i^{th} group (i.e. i = 1,…, m).

In other words, it is trying to minimize the confidence (weighting) factor α according to the following equation:

$$\alpha_{min} = max\ \frac{m - Q^2}{m - 1\ Q^2 + m}, 0 \tag{4.11}$$

ii) During Hierarchical Ensemble Kalman Filtering, the n ensemble members within each group are treated exactly as described for Plain EnKF except during the updating step where linear regression is involved. The regression computation proceeds as follows:

The regression coefficient, β_i , is computed for each of the n members within each of the i^{th} group, while the mean and standard deviation of each sample are computed, and thus the ratio Q.

The regression confidence factor α is computed from Eq. 4.11 above. The regression is completed for each ensemble by multiplying the sample regression coefficient β_i by the regression confidence factor. *Q* is defined as the following:

$$Q = \frac{\sqrt{\sum_{i=1}^{m}\left(\beta_i^{k,obs} - \overline{\beta}^{k,obs}\right)^2 \Big/ m-1}}{\overline{\beta}^{k,obs}} \tag{4.12}$$

where $\overline{\beta}^{k,obs} = \sum_{i=1}^{m} \beta_i^{k,obs} \Big/ m$ (4.13)

In the context of using Hierarchical EnKF on our synthetic field case, the regression factor $\beta_i^{k,obs}$ is the Kalman Gain at a particular grid block *k* for an ensemble member from the i^{th} group for a particular measurement *obs* (i.e. water-cut or bottom-hole pressure) at a certain data assimilation time. The fundamental goal of Hierarchical EnKF is to minimize the difference of the Kalman Gain at a grid block between the members from different groups at a certain data assimilation time. Results of implementing Hierarchical EnKF history matching on the 9-Spot synthetic field case are discussed in the following section.

Hierarchical Ensemble Kalman Filter Assimilation Results

Figure 4.15 shows the map of the multiplier values for each well between bottom-hole pressure and log permeability after Hierarchical EnKf history matching at *t=4000days*.

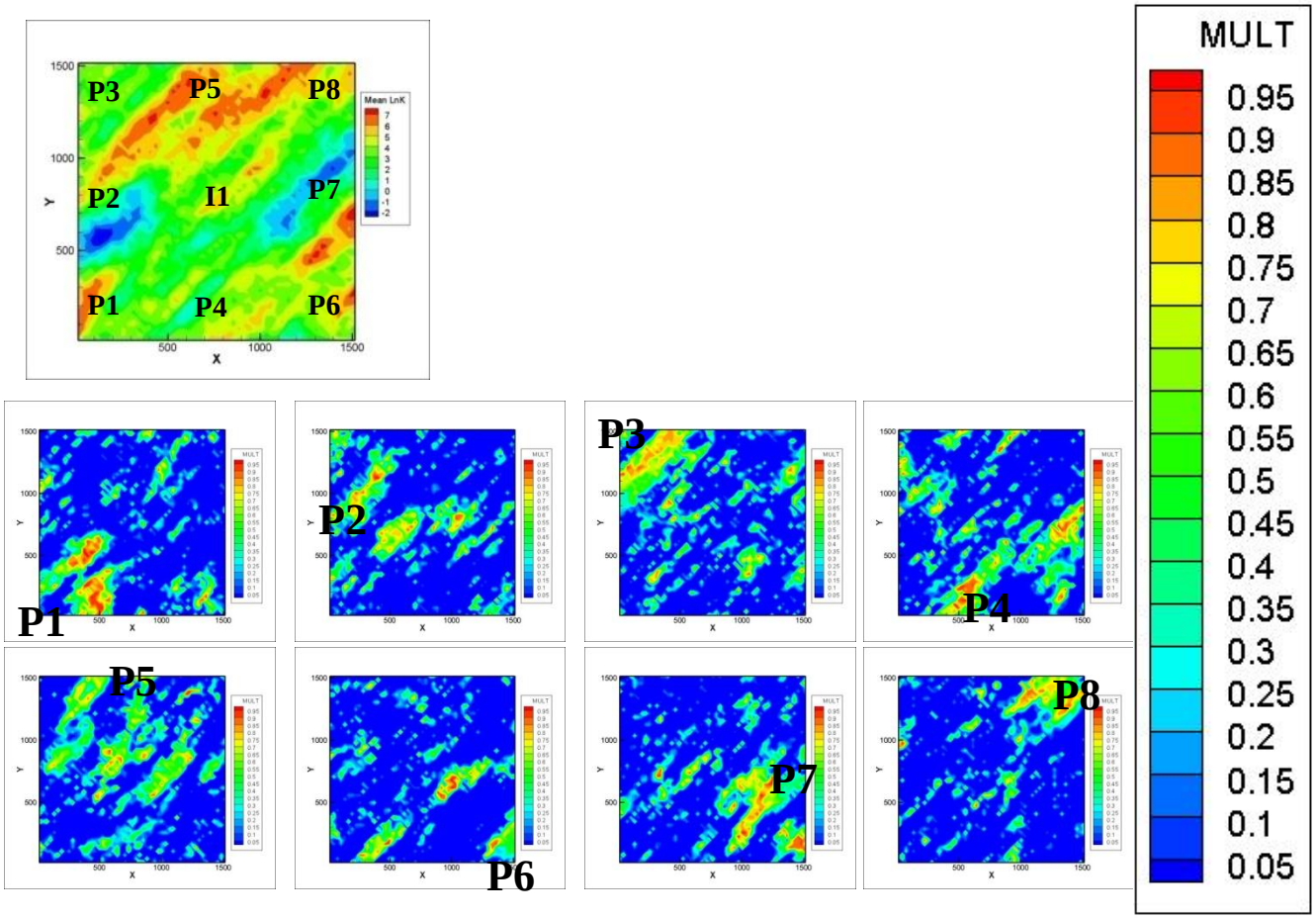


Figure 4.15. The map of the multiplier values for each well between bottom-hole pressure and log permeability after Hierarchical EnKF history matching at *t=4000days*. The number on each picture dictates the specific well at which the cross-covariance between the model parameter (permeability) of the surrounding grid blocks and the observation data (water-cut) at that well is studied.

In this study, we divide the 160 initial realizations into 4 groups of 40 members. Then we implement the Hierarchical EnKF algorithm and run the forward simulation using one group of 40 final updated members. The pictures in the Figure 4.16 below show the simulated dynamic responses of these 40 selected members after Hierarchical EnKF history matching as well as those of Plain EnKF with an ensemble size of 40. We have a better history match with Hierarchical EnKF, with the reference model responses being captured within the spread of ensemble responses, as well as from an RMS minimization point of view.

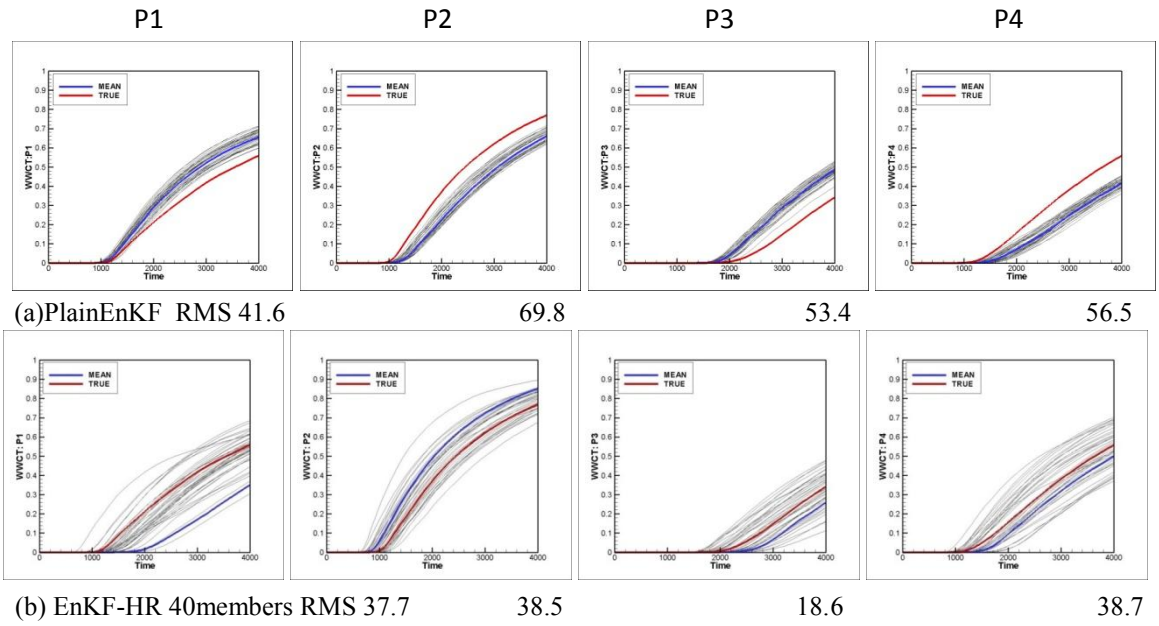


Figure 4.16. The EnKF WWCT matching results of selected producers: P1, P2, P3, and P4 with: (a) Plain EnKF of 40 members, (b) Hierarchical EnKF with 4 X 40 members with a group of 40 updated members simulated. The number under each plot is the root mean square (RMS) value.

In order to investigate the effect of using a different group size (but retaining the same initial number of emsemble members) during Hierarchical EnKF, the following cases are implemented on the same 9-Spot synthetic field case but with an assimilation frequency at 200 day-interval instead of 100-day interval (used in the previous cases). The cases studied are:

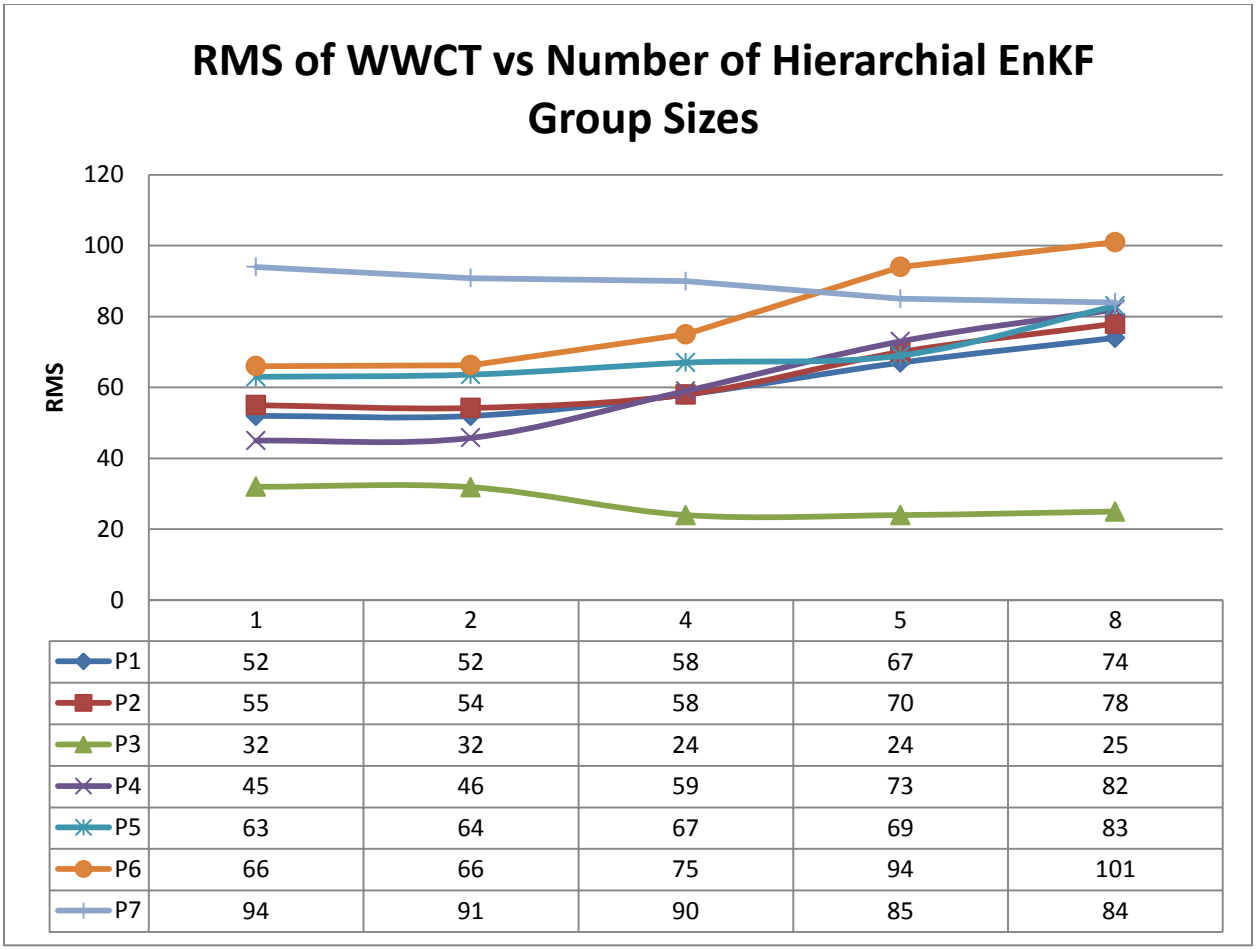
- 1 group of 160 members,
- 4 groups of 40 members,
- 5 groups of 36 members, and
- 8 groups of 20 members.

The RMS values of water-cut responses for different Hierarchical EnKF group sizes are shown in the following Table 4.3. The red number(s) in respect to each producer show the lowest RMS value(s) owned by the specific case among all cases. It was expected that the case of 1 group of 160 members would show the best history matching results since this case utilize all 160 members.

However, results in the following table show that group sizes of 1 and 2 are tied in having the most numbers of wells (four wells) with the lowest RMS values. In other words, the history matching quality is not degraded despite the use of less members in each group (i.e. 2 groups of 80 members as opposed to 1 group of 160 members) using Hierarchical EnKF. These results suggest that not necessarily the smallest group size/largest member size but a certain intermediate pick can optimize the application of Hierarchical EnKF.

Table 4.3: RMS values of water-cut response for different Hierarchical EnKF groups sizes

Case	P1	P2	P3	P4	P5	P6	P7	P8
1X160	52	55	32	45	63	66	94	128
2X80	52	54	32	46	64	66	91	125
4X40	58	58	24	59	67	75	90	214
5X36	67	70	24	73	69	94	85	228
8X20	74	78	25	82	83	101	84	284



	1	2	4	5	8
P1	52	52	58	67	74
P2	55	54	58	70	78
P3	32	32	24	24	25
P4	45	46	59	73	82
P5	63	64	67	69	83
P6	66	66	75	94	101
P7	94	91	90	85	84

Figure 4.17. RMS values of the water-cut responses vs the different Hierarchical EnKF group sizes: 1, 2, 4, 5 and 8 (x-axis). Group size of 1 refers to the case of 1 group of 160 members, 2 refers to 2 groups of 80 members, etc.

According to the above Figure 4.17, different group sizes show variant water-cut history matching quality, from an RMS minimization point of view. It shows that sensitivity of the assimilation quality to different group sizes needs to be addressed for the optimum application of Hierarchical EnKF.

Results of implementing the various covariance localization schemes: Distance-based, Streamline trajectory-based, Streamline sensitivity-based, along with Plain EnKF on the 9-Spot synthetic field case will be discussed in the following section.

Comparison of History Matching Results among Covariance Localization Schemes

Referring to Figure 4.18 and Figure 4.19, among all cases, Case (b) Distance based covariance localization has the lowest RMS values for four wells: P2, P3, P4, and P5. Case (c) EnKF with Streamline trajectory based covariance localization has the lowest RMS values for two wells: P1 and P8. Case (a) Plain EnKF has the lowest RMS values for two wells: P2 and P3. In addition, the water-cut history match results of Plain EnKF have the lowest RMS values for wells P6 and P7.

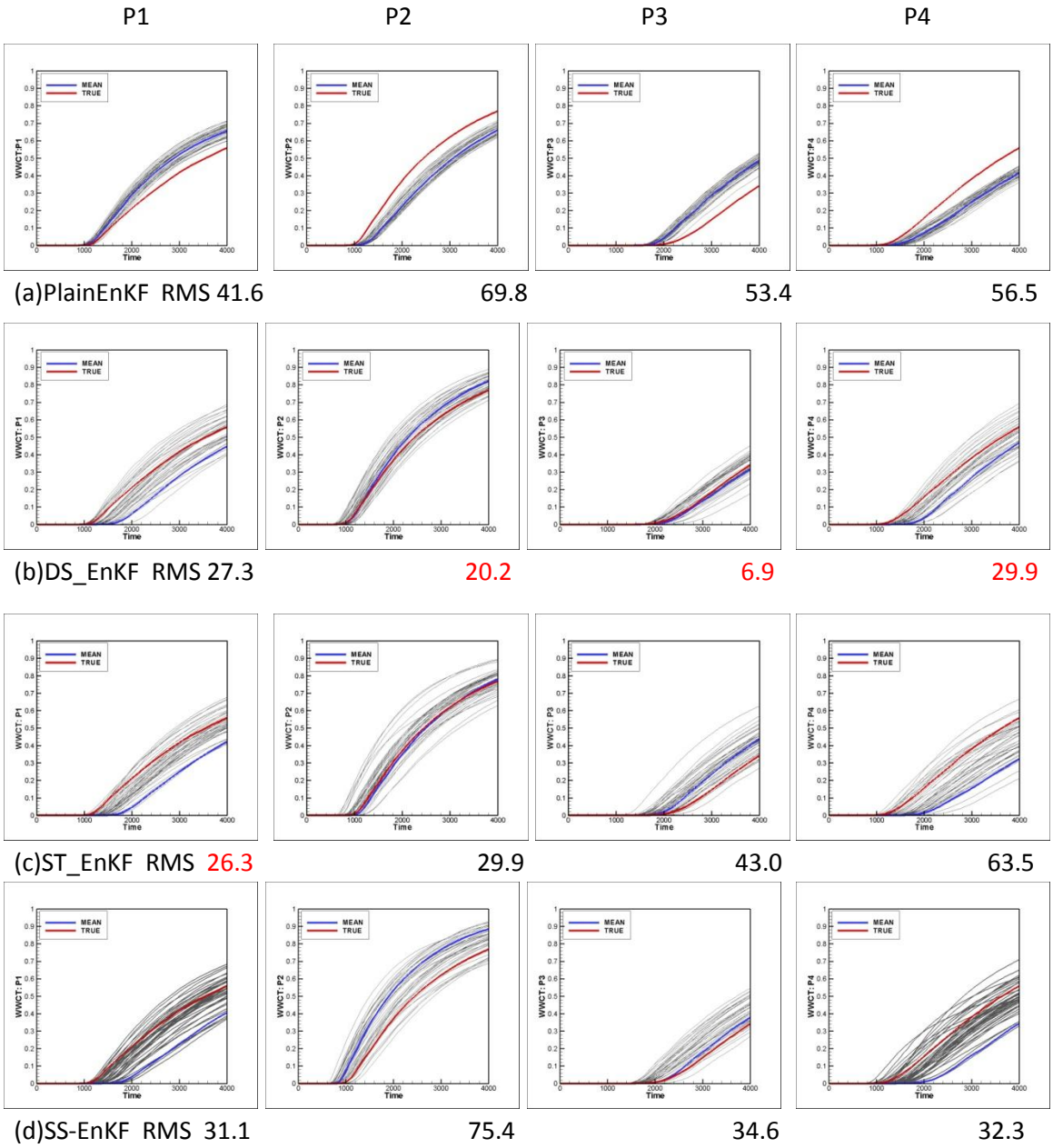


Figure 4.18. The EnKF water-cut matching results of each of the four producers: P1, P2, P3 and P4 for the case with (a) Plain EnKF, (b) EnKF with Distance based covariance localization, (c) EnKF with Streamline trajectory based covariance localization, and (d) EnKF with Streamline sensitivity based covariance localization. Red line dictates reference model response while blue line dictates mean of the ensemble responses. The number under each plot is the root mean square (RMS) value. The red numbers indicate, for each well, the lowest RMS values held by the particular localization scheme among all cases.

In summary, Distance based covariance localization (when the optimal cut-off limit of 900 is used) has the most number of producers with the lowest RMS values.

However, referring to wells P6 and P7, all localization approaches cases (b) through (d) overall have relatively large spread to capture the reference model response, despite their higher RMS values than those of Plain EnKF. Overall, even for other producers, all cases with covariance localization have wider spread of the ensemble responses compared to those of Plain EnKF.

In the following section, bottom-hole history matching results for these localization schemes implemented on producers P1, P2, and P3 among the nine wells will be shown.

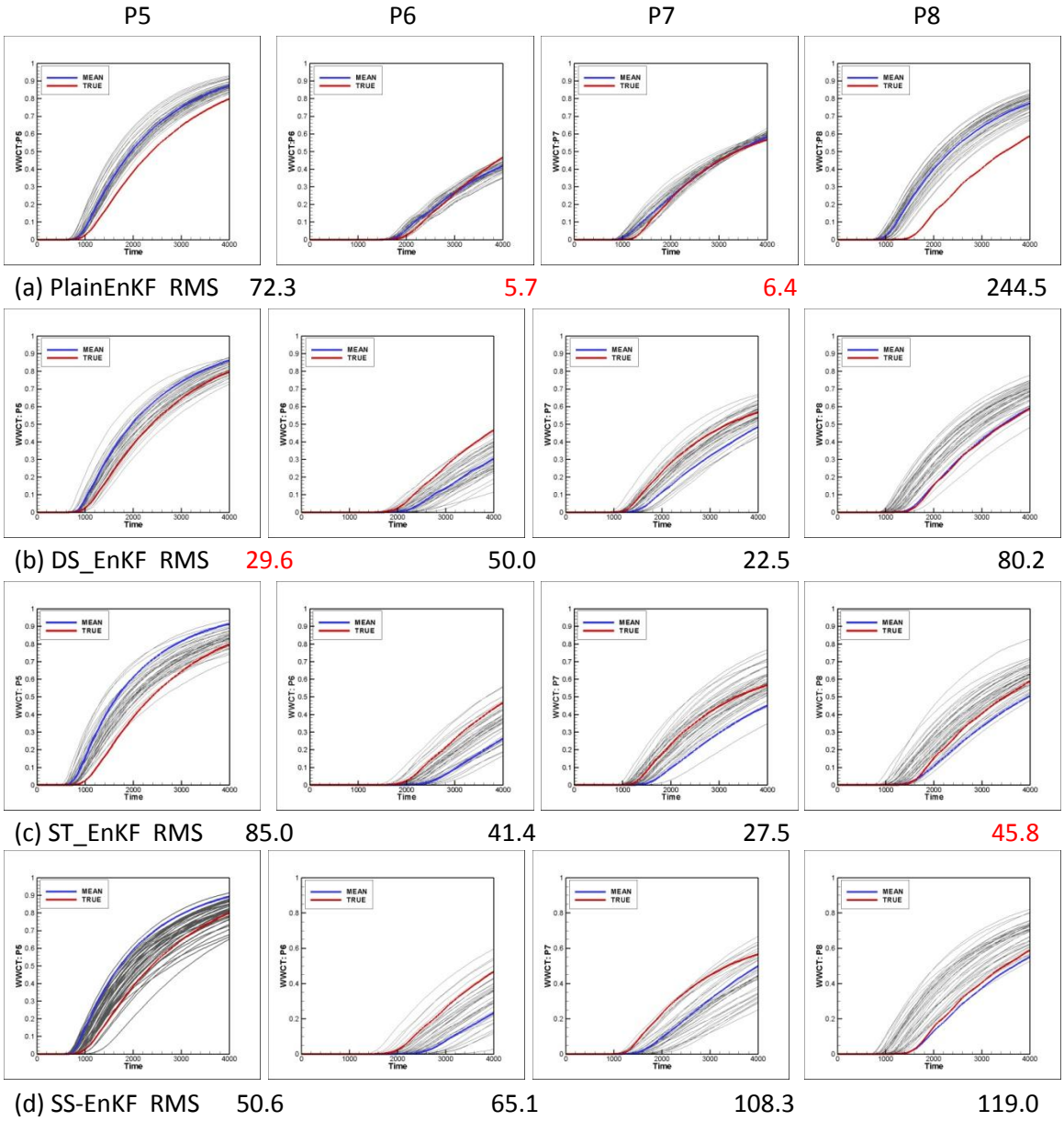


Figure 4.19. The EnKF WWCT matching results of each of the four producers: P5, P6, P7 and P8 for the case with (a) Plain EnKF, (b) EnKF with Distance based covariance localization, (c) EnKF with Streamline trajectory based covariance localization, and (d) EnKF with Streamline sensitivity based covariance localization. Red line dictates reference model response while blue line dictates mean of the ensemble responses. The number under each plot is the root mean square (RMS) value. The red numbers indicate, for each well, the lowest RMS values held by the particular localization scheme among all cases.

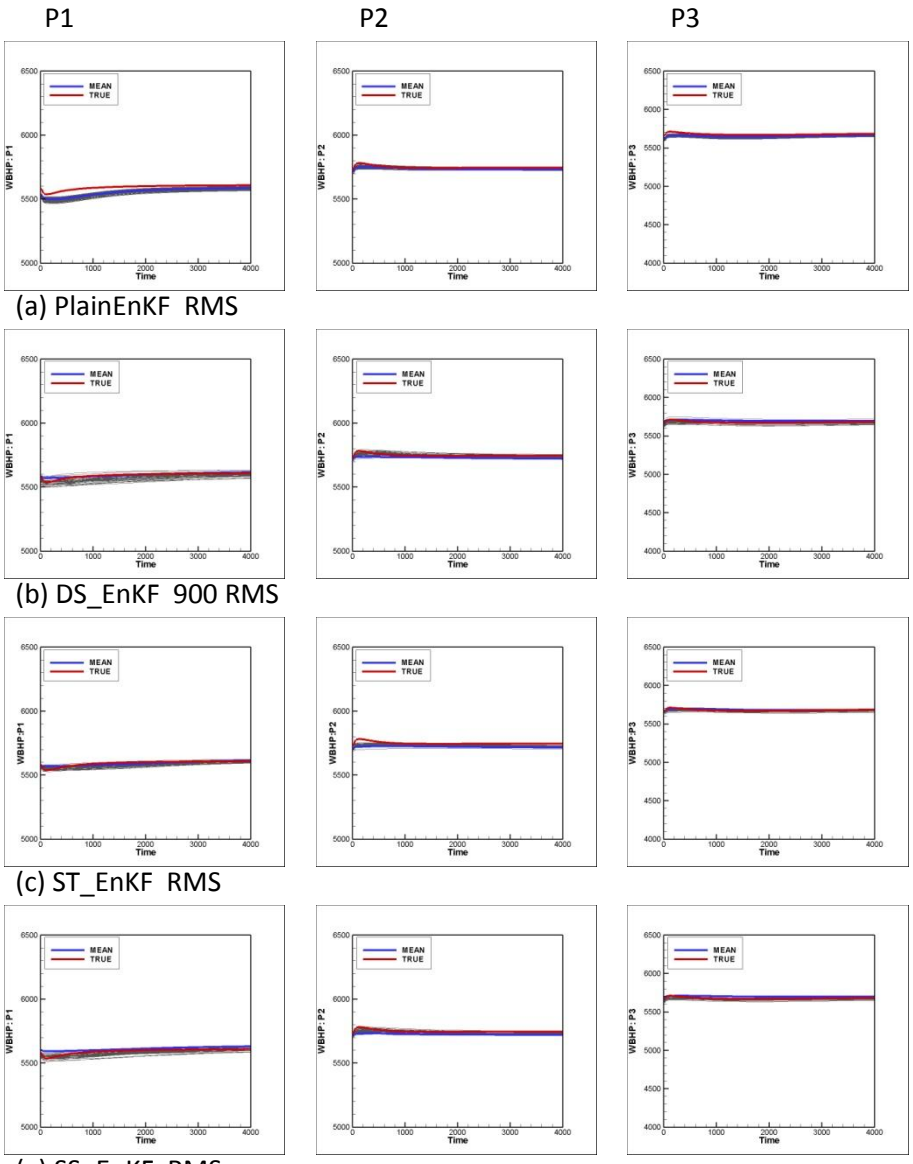
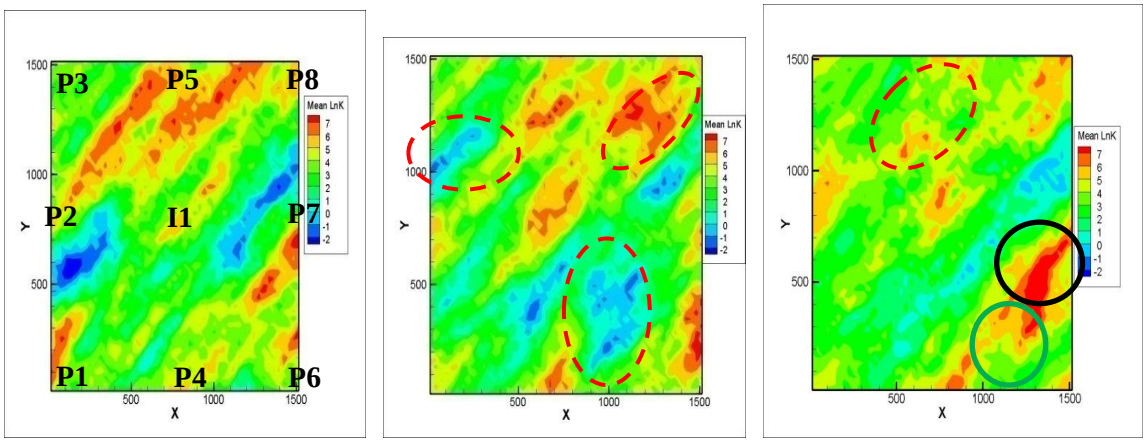
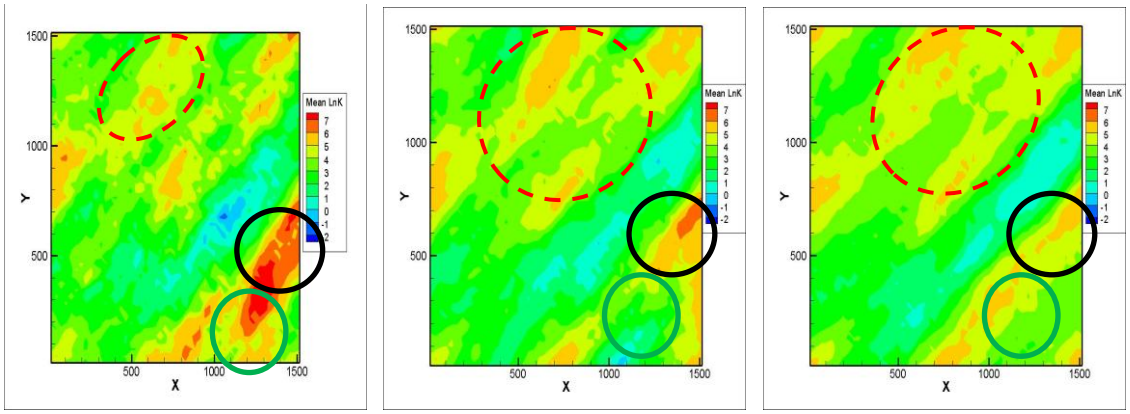


Figure 4.20. The EnKF WBHP matching results of three selected wells: P1, P2 and P3 for the cases with (a) Plain EnKF, (b) EnKF with distance based covariance localization, (c) EnKF with streamline trajectory based covariance localization, and (d) EnKF with streamline sensitivity based covariance localization.

The above Figure 4.20 shows that the bottom-hole pressure matching results for producers P1, P2, and P3 for the covariance localization schemes. Overall, all cases have comparable bottom-hole pressure profile matches.



(a) Reference Model Permeability (b) Plain EnKF Permeability (c) EnKF-ST Permeability



(d) EnKF-SS Permeability (e) EnKF-DS Permeability (f) EnKF-HR Permeability

Figure 4.21. Mean of the final updated ensemble permeability fields for the localization study; (a) Reference model, (b) Final updated mean of ensemble permeability for EnKF with no localization, (c) Final updated mean of ensemble permeability for EnKF with streamline trajectory based covariance localization (EnKF-ST), (d) Final updated mean of ensemble permeability for EnKF with streamline sensitivity based covariance localization (EnKF-SS), (e) Final updated mean of ensemble permeability for EnKF with distance based covariance localization (EnKF_DS), and (f) Final updated mean of ensemble permeability for hierarchical based EnKF (EnKF-HR).

In terms of the quality of parameter estimation, the final updated permeability fields of all cases are demonstrated in the above Figure 4.21. Without any covariance localization, the permeability undershoots and overshoots are apparent in Plain EnKF [case (b)] as seen in the three highlighted regions. However, when various kinds of covariance localization is

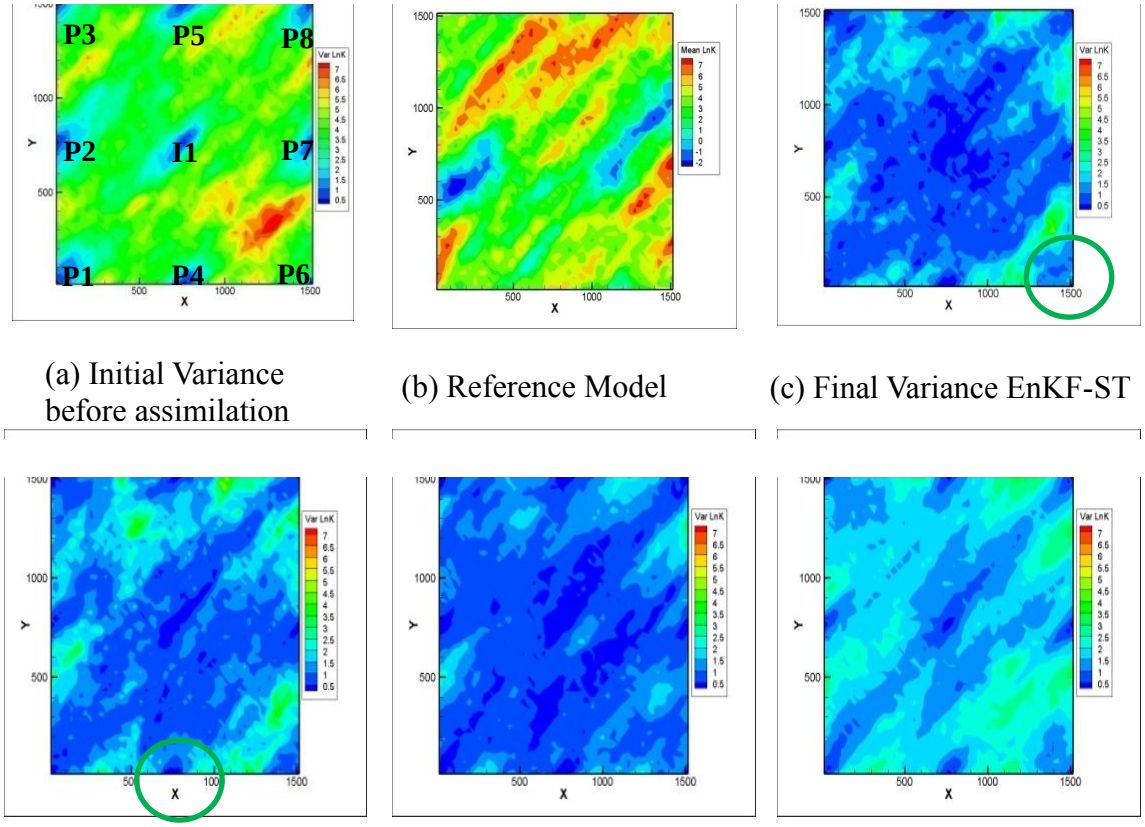
implemented [cases (c) through (f)], the undershooting/overshooting in these areas appears mitigated.

With the aid of Streamline-based covariance localization [cases (c) and (d)], both cases are able to capture the geological information of the reference model, as shown in the ellipsoid region highlighted with black solid lines in the bottom right corner of cases (c) and (d) respectively.

However, comparing the Streamline based approaches to either EnKF aided with Distance-based covariance localization or Hierarchical EnKF, the permeability is overestimated highlighted as the green circled regions in the bottom right corner of cases (c) and (d) respectively.

Overall, Distance-based covariance localization and Hierarchical EnKF have smoother parameter estimation, whereas the streamline-assisted approaches are able to capture the high permeability channels, but at the expense of an overestimation of parameters at certain regions.

In the following section, the variance of the mean of the ensemble updated permeability with respect to the true (reference model) permeability, at the final assimilation time, for each covariance localization scheme, is illustrated.



(a) Initial Variance before assimilation (b) Reference Model (c) Final Variance EnKF-ST

(d) Final Variance EnKF-SS (e) Final Variance EnKF-DS (f) Final Variance EnKF-HR

Figure 4.22. Variance of the ensemble permeability fields for the localization study; (a) Initial variance, (b) Permeability distribution of the reference model, (c) Final variance for EnKF with Streamline trajectory based covariance localization, (d) Final variance for EnKF with Streamline sensitivity based covariance localization, (e) Final variance for EnKF with Distance based covariance localization, and (f) Final variance for Hierarchical based EnKF.

The equation used to compute the variance at each grid block is as follows:

$$\text{var } iance[\log(k)] = \frac{1}{M-1}\sum_{i=1}^{M}[\log(k_i) - \log(k_{ref})]^2 \tag{4.14}$$

where *log(k)* is the logarithm of the permeability for each of the ensemble members, *i*, while k_{ref} is the reference model permeability.

According to Figure 4.22, it is apparent for all cases with covariance localization cases (c) through (f) that the variance distribution is along the direction of SW-NE. This coincides with the direction of the reference model's permeability spatial distribution. The higher variance (light blue regions) as seen in cases (c) and (d) coincides with the high permeability streaks of the reference model. In other words, the grid blocks with higher permeability have the final updated permeability farther away from the true; the grid blocks with lower permeability have the final updated permeability closer to the true.

It is evident from cases (c) and (d) that the higher variance as circled are in agreement with the overshooting regions as seen in the earlier Figure 4.21.

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

Various covariance localization approaches used to improve the application of the Ensemble Kalman Filter in the history matching of petroleum reservoirs are proposed. The characteristics and methodology of each approach are demonstrated through its application on two main synthetic reservoir models (5-spot and 9-spot respectively). The quality of the history match (water-cut, bottom-hole pressure), parameter (log permeability) estimation, variance of final updated permeability compared to the reference model permeability, etc have all been investigated to determine the performance of each approach. Based on the results in this study, the following conclusions and recommendations are made.

Characteristics of the EnKF approach:

- Using the Kalman Filter, the update covariance matrix is computed explicitly. However, with the Ensemble Kalman Filter, the entire update covariance matrix is computed from the propagated ensemble. In other words, this covariance is a sample covariance and does not need to be computed explicitly (which is the case for Kalman Filter). The ensemble is typically orders of magnitude smaller than the number of state variables. The number of state variables is the product of number of grid blocks and number of state variable types. This enables the EnKF methodology to be more computationally efficient than the Kalman Filter.

- However, when a small ensemble size is used for a large, field-scale model, poor estimate of the cross-covariance could occur. To mitigate such problem, various covariance localization methods are used to improve EnKF performance, without the use of large ensemble sizes that require enormous computational resources.

Characteristics of the covariance localization approaches with respect to EnKF data assimilation on a 9-spot synthetic field case:

- Considering the quality of the parameter estimation, overall, all covariance localization schemes are able to mitigate the parameter overshooting and undershooting problems experienced in Plain EnKF applications.
- Both Distance-based covariance localization and Hierarchical EnKF have smoother parameter estimation; however, the estimation may be "too smooth" as certain high permeability channels have not been captured. However, the alternative Streamline-assisted approaches have been able to achieve this.
- Streamline-based approaches are able to capture the high permeability channels/low permeability patches, but at the expense of permeability overestimation at certain regions.
- The Distance based covariance localization approach gives us good history matching quality, from an RMS minimization point of view, as long as the choice of the length scale is optimal. Although the selection of the length scale can be aided by the prior geological knowledge; this process requires trial and error.

- Compared to the history matching quality of Plain EnKF applications, all covariance localization methods have the ensemble response with wide enough spread that captures the reference model response, despite the higher RMS values with certain wells.
- Hierarchical EnKF results suggest that not necessarily the smallest group size/largest member size but a certain intermediate pick can optimize the application of Hierarchical EnKF.

Recommendations

- Sensitivity of the assimilation quality to the following needs to be addressed or defined for the optimum implementation of EnKF:
  - the total assimilation time and frequency of the assimilation steps;
  - the different types of observation data (water-cut, bottom-hole pressure, etc) to be assimilated;
  - the combinations of the measurement errors of each of the observation types.
- Further research can be performed to investigate the possible reasons for the overshooting problems experienced during Streamline-based covariance localization. This may be due to the assumption of linearity between permeability and observed data at those regions.
- Sensitivity of the assimilation quality to different group sizes should be addressed for the optimum application of Hierarchical EnKF.

REFERENCES

Alpak, F. O., Barton, M.D., and Caers, J., A Flow-based Pattern Recognition Algorithm for Rapid Quantification of Geologic Uncertainty, *Computational Geosciences*, **14**(4), 603-621, 2008.

Anderson, J. L., A Hierarchical Ensemble Filter for Data Assimilation, *Monthly Weather Review*, **131**, 634, 2003.

Anderson, J. L., and Anderson, S.L., A Monte-Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts, *Monthly Weather Review,* **127**(12): 2741-2758, 1999.

Arroyo, E., Devegowda, D., and Datta-Gupta, A., Streamline-Assisted Ensemble Kalman Filter for Rapid and Continuous Reservoir Model Updating, *SPE Reservoir Evaluation & Engineering,* **11**(6): 1046-1060, 2006.

Borg, I., and Groenen, P, *Modern Multidimensional Scaling: Theory and Application*, 2nd ed., Springer-Verlag, New York, 2005.

Devegowda, D., Arroyo-Negrate, E., and Datta-Gupta, A., Flow Relevant Covariance Localization During Dynamic Data Assimilation Using EnKF, *Advances in Water Resources,* **33**: 129-145, 2009.

Emanuel, A. S., and Milliken, W.J., History Matching Finite Difference Models with 3D Streamlines, paper SPE 49000 presented at SPE Annual Technical Conference and Exhibition, New Orleans, Louisiana, September, 27-30, 1998.

Evensen, G, *Data Assimilation: The Ensemble Kalman Filter*, Springer, New York, 2006.

Furrer, R., and Bengtsson, T., Estimation of High-Dimensional Prior and Posterior Covariance Matrices in Kalman Filter Variants, *J. of Multivariate Analysis,* **98**: 227-255, 2004.

Gaspari, G., and Cohn, S.E., Construction of Correlation Functions in Two and Three Dimensions, *Office Note Series on Global Modeling and Data Assimilation,* DAO Office Note 96-03R1, 1996.

Hamill, T. M., Whitaker, J.S., and Snyder, C., Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter, *Monthly Weather Review,* **129**: 2776-2790, 2001.

Jolliffe, I.T., *Principal Component Analysis*, Series: *Springer Series in Statistics*, 2nd ed., Springer, New York, 2002.

Ng, A.Y., Jordan, M.I., and Weiss, Y., On Spectral Clustering: Analysis and an Algorithm, *Neural Information Processing Systems (NIPS)* ***14***, 2002.

Scheidt, C. and Caers, J., Using Distances and Kernels to Parameterize Spatial Uncertainty for Flow Applications, EAGE Petroleum Geostatistics Conference. Cascais, Portugal, 10-14 September, 2007.

Spartek Systems Geophysical Instrumentation, SPSRO^{TM} Permant Monitoring. from http://www.sparteksystems.com/siteimages/Brochures/Flyer_SPSRO_Permanent_Monitoring_20090815.pdf, 2008.

VITA

Name: Yeung Yip

Address: 3116 Mail Drop, Texas A&M University, TX 77843

Email Address: yeung.yip@pe.tamu.edu

Education: B.S., Petroleum Engineering, Texas A&M University, 2008

M.S., Petroleum Engineering, Texas A&M University, 2011